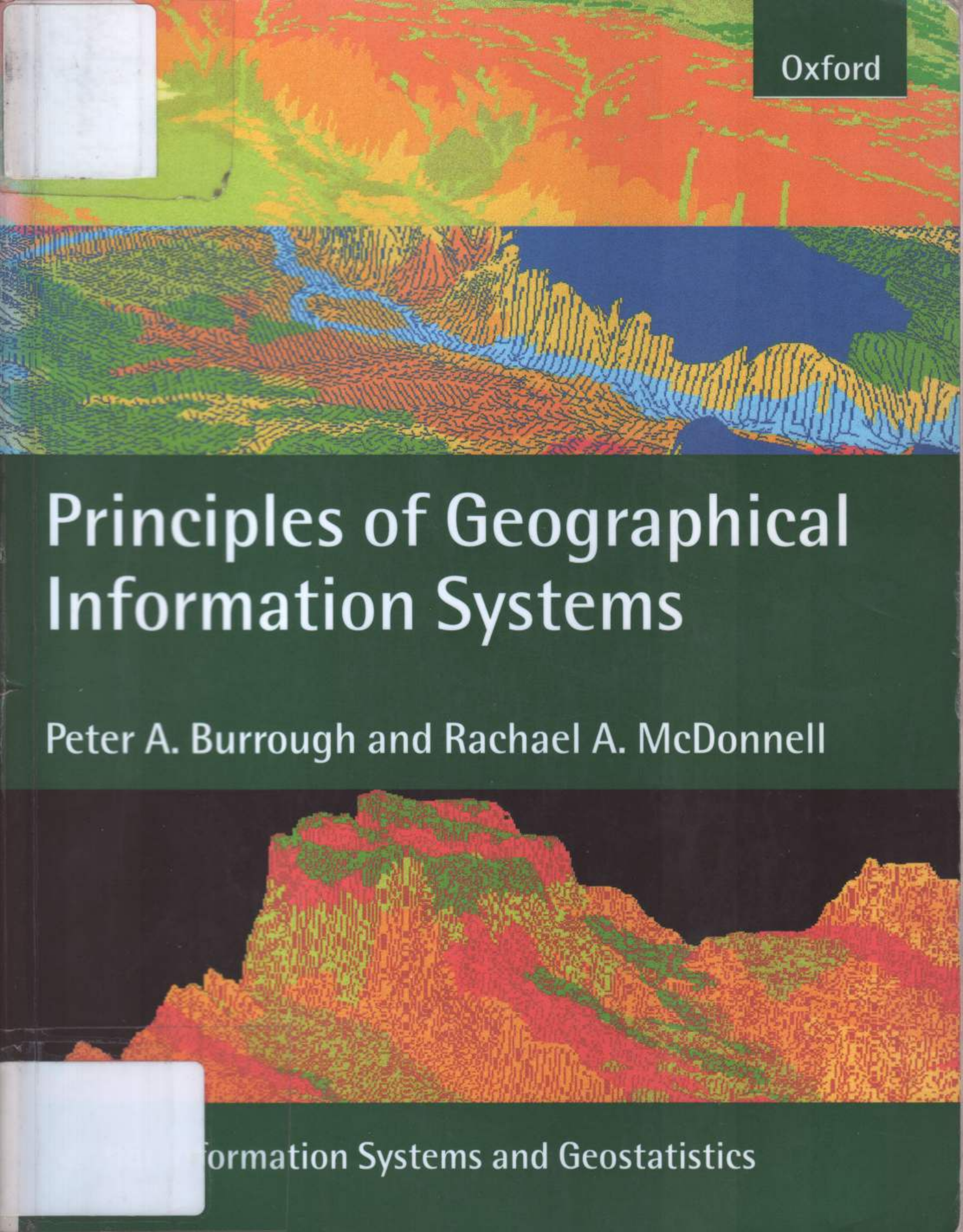


Oxford
Principles of Geographical Information Systems
Peter A. Burrough and Rachael A. McDonnell
formation Systems and Geostatistics

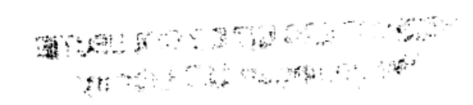
9950

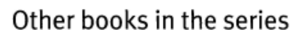
Principles of Geographical Information Systems

Spatial Information Systems and Geostatistics

General Editors

P. A. Burrough
M. F. Goodchild
R. A. McDonnell
P. Switzer
M. Worboys





Other books in the series

Anthropology, Space, and Geographic Information
M. Aldenderfer and H. D. G. Maschner

Spatial and Temporal Reasoning in Geographic Information Systems
M. J. Egenhofer and R. G. Golledge (eds.)

Environmental Modeling with GIS
M. Goodchild, B. O. Parks, L. Steyaert (eds.)

Managing Geographic Information Systems Projects
W. E. Huxold and A. G. Levinsohn

GIS County User Guide: Laboratory Exercises in Urban Geographic Information Systems
W. E. Huxold, P. S. Tierney, D. R. Turnpaugh, B. J. Maves, and K. T. Cassidy

Introduction to Disjunctive Kriging and Non-linear Geostatistics
J. Rivoirard

Principles of Geographical Information Systems

Peter A. Burrough

AND

Rachael A. McDonnell



OXFORD UNIVERSITY PRESS

1998

Oxford University Press, Great Clarendon Street, Oxford OX2 6DP
Oxford New York
Athens Auckland Bangkok Bogota Bombay Buenos Aires
Calcutta Cape Town Dar es Salaam Delhi Florence Hong Kong Istanbul
Karachi Kuala Lumpur Madras Madrid Melbourne Mexico City
Nairobi Paris Singapore Taipei Tokyo Toronto Warsaw
and associated companies in
Berlin Ibadan

Oxford is a trade mark of Oxford University Press

Published in the United States
by Oxford University Press Inc., New York

First published 1998
Reprinted with corrections 1998

British Library Cataloguing in Publication Data
Data available

Library of Congress Cataloging in Publication Data
Burrough, P. A.
Principles of geographical information systems / Peter A. Burrough and Rachael A. McDonnell.
p. cm. — (Spatial information systems)
Rev. edn. of Principles of geographical information systems for land resources assessment.
Includes bibliographical references and index.
1. Geographic information systems. I. McDonnell, Rachael.
II. Burrough, P. A. Principles of geographical information systems for land resources assessment. III. Title. IV. Series.
G70.212.B87 1997 910′.285—dc21 97–25863

ISBN 0–19–823366–3
ISBN 0–19–823365–5 (Pbk)

10 9 8 7 6 5 4 3 2

Printed in Great Britain
on acid-free paper by Butler & Tanner Ltd, Frome, Somerset

To Joy, Alexander, Nicholas, and Gerard

Preface

As with the first edition (Burrough 1986), this book describes and explains the theoretical and practical principles for handling spatial data in geographical information (GI) systems. In the mid-1980s these computer tools were being developed from computerized aided design packages and computer-aided manufacturing packages (CAD-CAM), from automated mapping systems, and from programs for handling data from remotely sensed images collected by sensors in satellites, and there was a need to provide a text that covered the common ground. Since then, sales of GIS as commercial products for the mapping and spatial sciences and for the inventory and management of spatial resources of cities, states, or businesses have increased phenomenonally. People have become fascinated by the power of immediate interaction with electronic representations of the world. The development of the cheap, powerful personal computer has facilitated this process and the current ability to link computers by electronic networks across the globe has provided a ready source of all kinds of data in electronic form, including maps and raster images taken from all kinds of platforms from satellites to hovercraft to deep-sea vehicles.

Today, not just maps are made using GIS, but the infrastructure of utilities (cables and pipes) in the streets of your town will be held in a GIS, your taxis and emergency services may be guided to their destinations using satellite-linked spatial systems, the range of goods in your shops may be decided by systems linking customer preferences to neighbourhood and socio-economic status, the foresters and farmers will be monitoring their stands and crops with spatial information systems, and a whole range of scientists and technicians will be advising governments, the military, and businesses from local to world scale how best to deal with the distribution of what interests them most.

The ready provision of powerful computing tools, the increasing abundance of spatial data, and the enormous diversity of applications that are not just limited to map-making or scientific enquiry, make it even more important today than in 1986 to understand the basic principles behind GIS. Everyone thinks they understand GIS because they think they understand maps. But to understand a map is to understand how a person once observed the world, how they formulated imperfect, but plausible models to represent their observations, and how they decided to code these models on paper using conventional semiology. All these processes depend strongly on culture, on age, on discipline, and on background. With GIS we go even further down a path in the labyrinth that leads from perception to presentation of spatial information, because though computers can mimic human draughtsmen, they can do so only according to the ways they are programmed. The basic tenets of spatial interactions needed to describe a given spatial process or phenomenon may or may not be shared by the ground rules of the GIS you would like to buy. But electronic GIS provide an enormous wealth of choice not possible with conventional mapping. Whereas with conventional mapping (and even the earliest digital mapping) the map *was* the database, today the map is merely an evanescent projection of a particular view of a spatial database at a given time. On the one hand this gives us enormous power to review an unlimited number of alternatives and to make maps of anything, anyway we like, including dynamic modelling of spatial and temporal processes; on the other hand it may leave us swimming in the dark.

This book aims to provide an introduction to the theoretical and technical principles that need to be understood to work effectively and critically with GIS. It is based on the 1986 volume, but is much more than a simple second edition. The text examines the different ways spatial data are perceived, modelled conceptually, and represented.

It explains how, using the standard 'entity' and 'continuous field' approaches, spatial data are collected, stored, represented in the computer, retrieved, analysed, and displayed. As with the first edition, this material is accompanied by a critical evaluation of the sources, roles, and treatment of errors and uncertainties in spatial data, and their possible effects on the conclusions to be drawn from analyses carried out with GIS. Also included is a discussion of the principles and methods of dealing with uncertainty in spatial data, either through the probabilistic medium of geostatistics, or the possibilistic route using fuzzy logic. The principles of the latter are juxtaposed with the crisp concepts underlying the entity and field models so commonly used. Frequently encountered, specialist terms are explained in the glossary.

Spatial analysis and GIS are nothing without computer software, and the material in this book could not have been produced without recourse to many different programmes. Large, commercial systems may be unsuitable or unusable for teaching or scientific enquiry, and in Appendix 2 we provide information about sources of cheap or free software that have been used in preparing this book and which may be used to support training courses and research.

GIS is truly interdisciplinary and multidisciplinary, and the literature is spread over a range of journals. Trade journals such as *GIS World*, and *GIS Europe*, provide monthly coverage of technical and organizational developments; trade shows and conferences and scientific journals such as the *International Journal of Geographical Information Systems, Computers and Geosciences, Computers, Environment and Urban Systems, Geoinformation*, and *Transactions in GIS*, cover the core of the field, though much is published in discipline-specific journals, and also in 'grey' literature reports from conferences, institutes, and commerce. We have not covered everything in this volume: that is impossible, but we aim to provide a basic technical and scientific introduction to an important scientific and business activity, so that readers will better understand how GIS are being used to transform the ways in which much of our world is perceived, recorded, understood, and organized.

P.A.B. & R.A.McD.

Utrecht
February 1997

Acknowledgements

We thank the following persons for supplying Plates and material for figures and the permission to reproduce them: Ir Fred Hageman and Eurosense Belfotop NV, Wemmel, Belgium for Plates 1.1–1.6; Rotterdam Municipality for Plate 2.1; Dr Jan Ritsema van Eck (University of Utrecht) for Plate 2.2 and Figures 7.12 and 7.13; Dr Stan Geertman (University of Utrecht) for Plates 2.3 and 2.4; Mr Richard Webber of CCN Marketing for Plates 2.5–2.8; Drs Lodewijk Hazelhoff (University of Utrecht) for Plates 3.1–3.6; the Australian Environment Data Centre, Canberra for Plate 4.1; Dr Jan Clevers and the Agricultural University Wageningen, for Plate 4.2; Prof. Ulf Helldén, Lund University for Plate 4.3; Dr Robert MacMillan and The Alberta Research Council for Plate 4.4; and Dr Gerard Heuvelink for Plate 4.6; Prof. Tom Poiker for Figure 5.14; California Institute of Technology for Figure 1.2 (via Internet).

For base data used in examples and analyses we thank Prof. Andrew Skidmore (International Institute for Aerospace Survey and Earth Sciences (ITC), Enschede), Dr Brenda Howard (Institute of Terrestrial Ecology, Merlewood, Cumbria, UK), Prof. Boris Prister (Ukraine Institute of Agricultural Radiology, Kiev), Dr Robert MacMillan (formerly Alberta Research Council), Dr Alejandro Mateos (University of Maracai, Venezuela), RPL Legis (formerly Wageningen Agricultural University), Dr Arnold Bregt (Winand Staring Centre, Wageningen, the Netherlands), and Dr Ad de Roo, Dr Steven de Jong, Dr Victor Jetten, Drs Ruud van Rijn, and Drs Marten Rikken (all University of Utrecht).

For computer software and assistance with figures and data we thank our colleagues at the Department of Physical Geography, Netherlands Research Institute for Geoecology, Faculty of Geographical Sciences, University of Utrecht: Drs Lodewijk Hazelhoff, Dr Victor Jetten, Dr Willem van Deursen, Ing. Cees Wesseling, Dr Marcel van der Perk, Drs Derk Jan Karssenberg, Dr Paulien van Gaans, and Dr Edzer Pebesma; also Dr Gerard Heuvelink (now University of Amsterdam). We thank Taylor and Francis Ltd for permission to reproduce figures 2.5, 2.6, 11.2, 11.9, 11.2, and 11.13. Nicholas Burrough helped sort the figures.

For information on developments in computer networking we thank Gerard McDonnell (Novell Inc), and Peter Woodsford (Laserscan) for information on commercial developments in Object Oriented GIS. We thank Kevin Jones (Macaulay Land Use Research Institute, Aberdeen) and Prof. B. K. Horn (Massachusetts Institute of Technology) for information on slope algorithms. For the many discussions on the theory of spatial reasoning we thank Profs. Helen Couclelis, Waldo Tobler, and Michael Goodchild (NCGIA, Santa Barbara) and Prof. Andrew Frank (Technical University, Vienna) and many other colleagues too numerous to name. We also thank the Faculty of Geographical Sciences of Utrecht University for technical, financial and moral support over a period of more than twelve years.

Contents

List of Plates

(between pages 146 and 147)



1.1 Digital orthophoto of the city of Antwerp, Belgium



1.2 Digital orthophoto of a city with added vector line data to indicate administrative boundaries and data for utility mapping



1.3 Digital orthophoto of semi-rural area with motorway including added vector line data



1.4 Digital orthophoto draped over a digital elevation model to provide a perspective view of the landscape



1.5 Large-scale digital orthophoto of urban scene including overlaid vector data on roads and building outlines for a utilities database

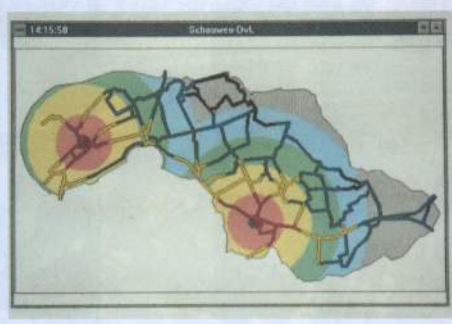


1.6 Vector data for the road and building outlines for Plate 1.5

Plates 1.1–1.6 reproduced by permission of Eurosense-Belfotop NV, Wemmel, Belgium

Plate 1. **Digital Orthophoto Maps and GIS**



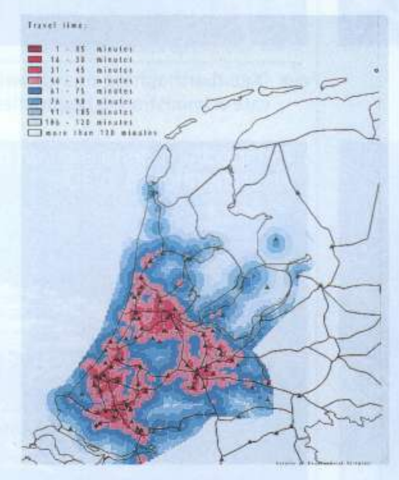


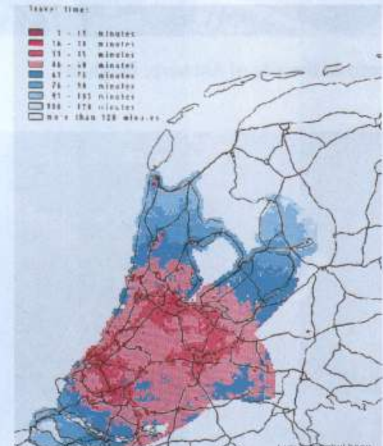
2.1 (right) Large-scale raster image of part of the digital database of Rotterdam showing utilities, cadastral information, and ownership details (in colours green and magenta)

Copyright/courtesy City of Rotterdam and the Computer Department, International Institute for Aerospace Survey and Earth sciences, Enschede)

2.2 (far right) A comparison of travel times via crow's flight distance and with the road

Courtesy Jan Ritsema van Eck





2.3 (right) Accessibility of places of work in the West of the Netherlands in terms of travel times by public transport

2.4 (far right) Accessibility of places of work in the West of the Netherlands in terms of travel times by private car

Plates 2.3 and 2.4 courtesy Stan Geertman





2.5 (right) The spatial prediction of marketing opportunities: the locations of Social Group A in Liverpool (red)

2.6 (far right) The spatial prediction of marketing opportunities: whisky consumption in Liverpool (red indicates households consuming most whisky)



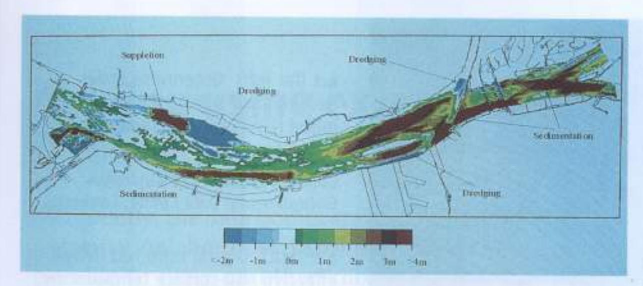


2.7 (right) The distribution of socio-economic group 'stylish singles' over the UK

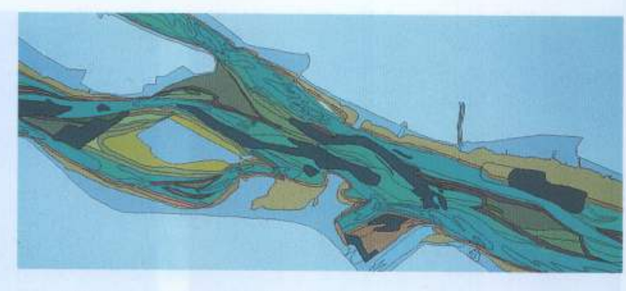
2.8 (far right) The distribution of the opportunities for selling packaged holidays in NE London

Plates 2.5–2.8 courtesy CCN Marketing

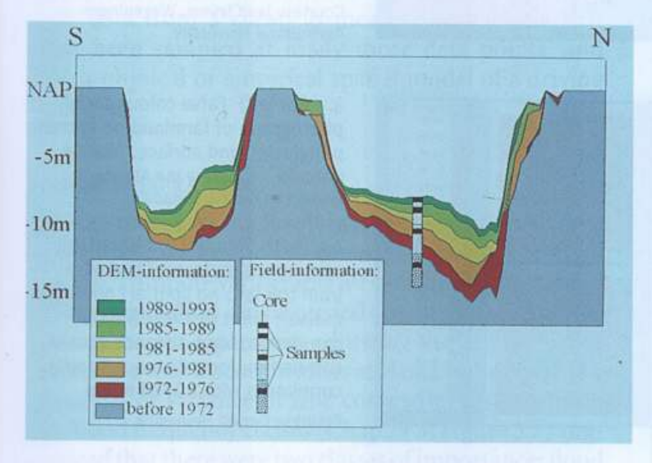
Plate 2. **Municipal planning, route finding, and marketing**



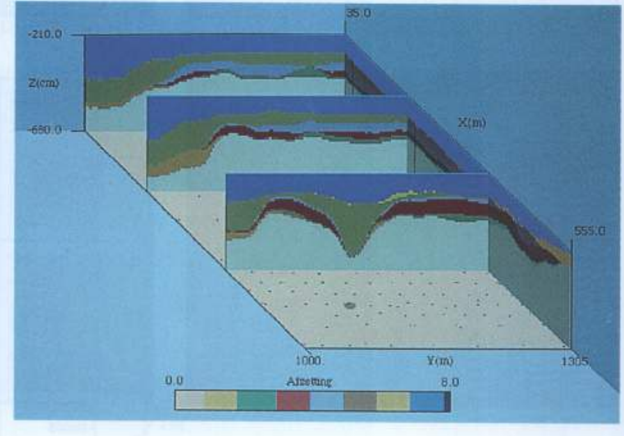
3.1 Changes in channel depths in the mouth of the river Rhine



3.2 The Geomorphological distribution of sediments and channels in the mouth of the rIver Rhine

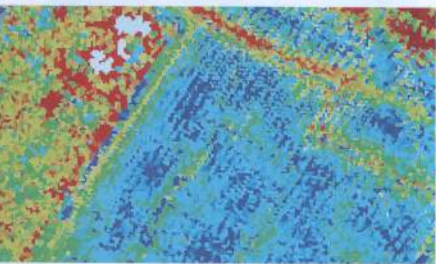


3.3. Cross-section through sediments of the rIver Rhine



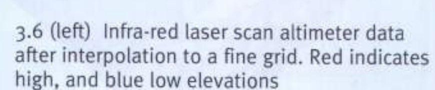
3.4. An example of fence-mapping to show continuous variation of sediment composition in three dimensions



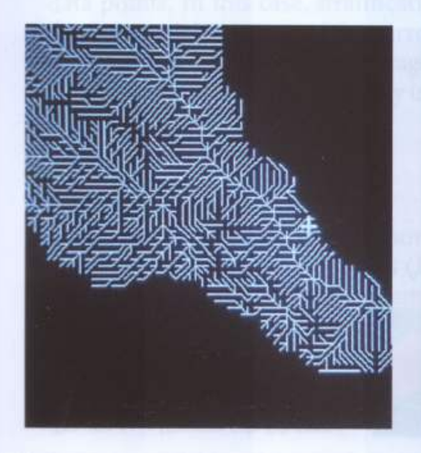


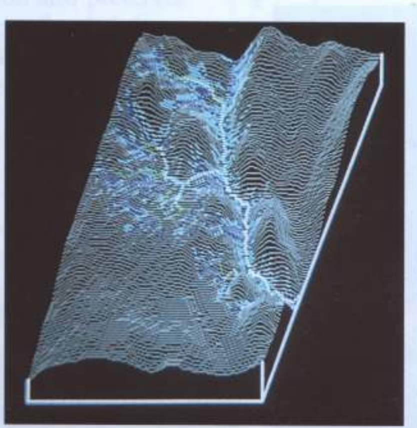
3.5 (far left) The distribution of sampling points of an airborne infra-red laser altimeter scanner

3.6 (left) Infra-red laser scan altimeter data after interpolation to a fine grid. Red indicates high, and blue low elevations



Plates 3.1–3.6 courtesy Lodewijk Hazelhoff and the Ministry of Public Works, The Netherlands



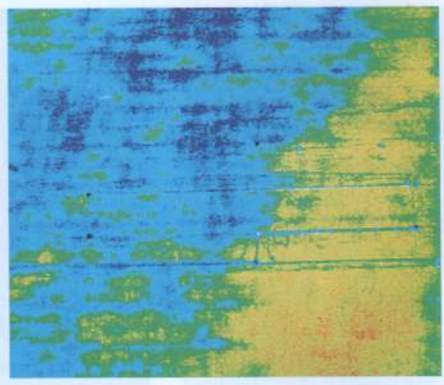


3.7 (far left) Example of local drain direction map created using the D8 algorithm

3.8 (left) Cumulative upstream elements (log scale) draped over the DEM from which they were derived to provide a perspective view of the drainage network

Plates 3.7–3.8 courtesy P. A. Burrough





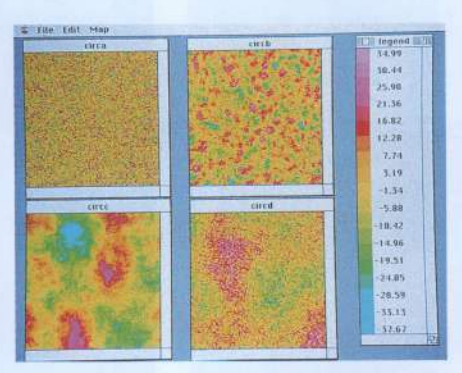
4.1 (far left) Greenness index (Normalized Difference Vegetation Index NDVI) for Australia, winter 1993)

Courtesy Internet site Australian Environmental Data Centre, Canberra

4.2 (left) Digital infra-red image of effective soil surface temperature, Minderhoudhoeve Experimental Farm, Dronten, The Netherlands

Courtesy Jan Clevers, Wageningen Agricultural University

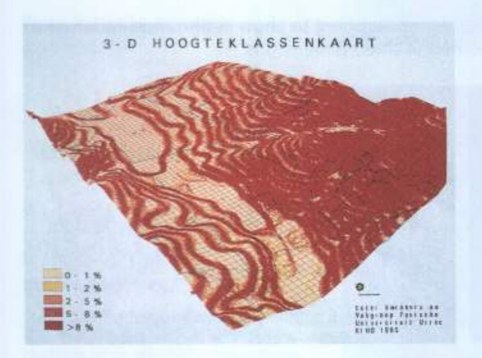


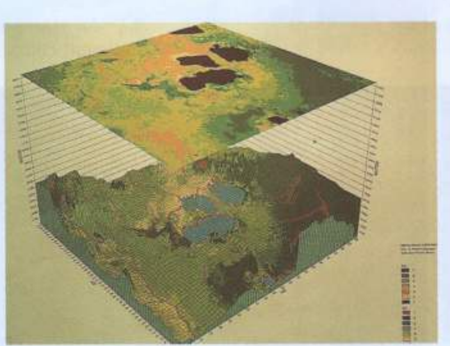


4.3 (far left) False colour aerial photograph of farmland on former periglacial land surface, Alberta, Canada Courtesy the Alberta Research Council, Canada

4.4 (left) Simulated, discretized continuous surfaces. Clockwise from top left: no spatial correlation (noise); short distance spatial correlation; long distance spatial correlation; long distance spatial correlation with added noise

Courtesy Gerard Heuvelink

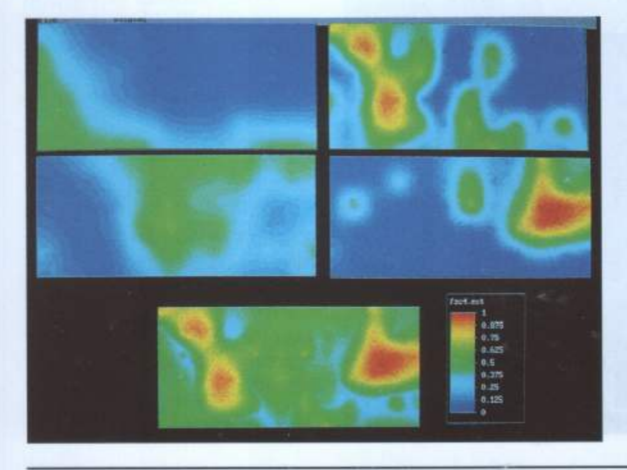




4.5 (far left) Tiger stripes of erroneously large slope angles obtained when an inappropriate interpolation method is used to generate a DEM from digitized contour lines

Courtesy Faculty of Geographical Sciences, Utrecht University

4.6 (left) Interpreted land cover data derived from Thematic Mapper satellite imagery draped over a DEM Courtesy Ulf Helldén, Lund University, Sweden





4.7 (far left) Top: Continuous mapping of four fuzzy soil classes. Bottom: Map showing the maximum membership value of all four classes

4.8 (left) Map of the maximum membership values for all four classes with superposition of zones where the classes are maximally indeterminate (boundaries)

Plates 4.7–4.8 Courtesy P. A. Burrough

Plate 4. **Remote sensing and errors**

ONE

Geographical Information: Society, Science, and Systems

Geographical information sciences: a brief history

The growing world population is seriously increasing demands on the earth's resources of land, air, water, and raw materials. In previous times excess population could migrate to more sparsely populated areas or numbers were reduced by plague or war; these options are no longer possible nor acceptable to civilized society at the end of the twentieth century. At the same time, human societies are becoming more organized, not just to ensure that people have sufficient land and natural resources for basic needs, but to support the multifarious activities of increasingly complex social and economic behaviour patterns. As pressure on natural resources and land increases the greater is the need for properly organized agreements about how they should be shared not only for the benefit of humans, but for all forms of life. This requires not only an understanding of the spatial and temporal patterns of resources but also insight into the spatial and temporal processes governing their availability. History has demonstrated many times that the deterioration of renewable resources and the reduction of sustainable means of livelihood may produce tensions and stresses of overpopulation and pollution that increase to the point where civilized life breaks down. Aggressive confrontations between different groups of people over land occupation and land resources occur daily all over the globe.

People do not always resort to violence to solve disputes about land. Once humans had developed the basic notions of counting and arithmetic to ensure fairness in the trade and inheritance of animals, crops, and valuable objects (Barrow 1992, McLeish 1992) they were able to apply similar concepts to the mensuration and apportioning of land; this is particularly so in Western cultures. The codes and laws that have been developed for regulating and enforcing agreements for the division and use of land range from temporary stick signs set up in the rain forest or for mining claims to modern cadastral systems that are enshrined in national laws (Aldenderfer and Maschner 1996, Burrough 1975, Dale and McLaughlin 1988, De Soto 1993).

The development of laws and codes about land use and land ownership brought with it the need to establish records of transactions and agreements that were independent of individual or collective human memories. From the earliest civilizations to modern times spatial data have been collected by navigators, geographers, and surveyors to be recorded in a coded, pictorial form by map-makers and cartographers.

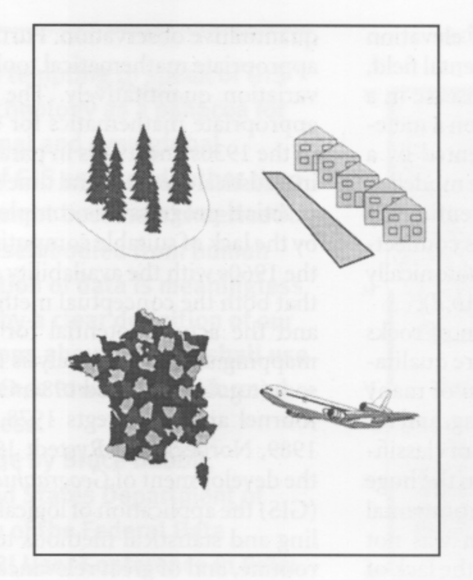
In Roman times, the *agrimensores*, or land surveyors, were an important part of the government, and the results of their work may still be seen in vestigial form in the European landscapes to this day (Dilke 1971).

The decline of the Roman Empire led to the decline of surveying and map-making which revived with the geographical discoveries of the Renaissance. By the seventeenth century skilled cartographers such as Mercator had demonstrated that not only did the use of a mathematical projection system and an accurate set of coordinates improve the reliability of the measurement and location of areas of land, but the registration of spatial phenomena through an agreed standard provided a model of the distribution of natural phenomena and human settlements that was invaluable for navigation, route finding, and military strategy (Hodgkiss 1981). By the eighteenth century the European states had reached a state of organization when many governments realized the value of systematic mapping of their lands. The *Geographical Information Society* was first created through the establishment of national government bodies whose mandate was to produce cadastral and topographical maps of whole countries. These highly disciplined institutes have continued to this day to render the spatial distribution of the features of the earth's surface, or topography, into map form. During the last 200 years many individual styles of map have been developed, but there has been a long, unbroken tradition of high cartographic standards that has continued until the present.

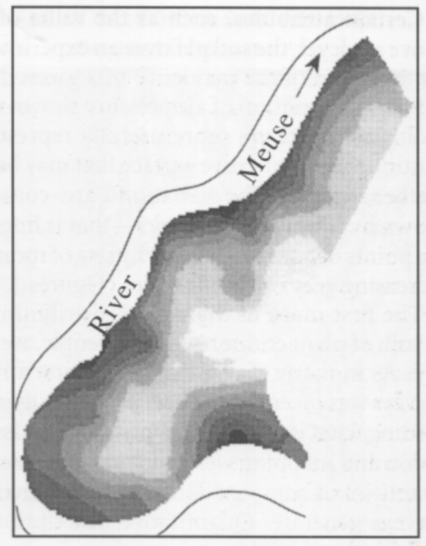
The mapping sciences—geodetical surveying, photogrammetry, and cartography—developed a powerful range of tools for accurately recording and representing the location and characteristics (usually referred to as attributes) of well-defined natural and anthropomorphic phenomena (e.g. Goudie *et al.* 1988, Johnston *et al.* 1988). The basic units of geographic information were decided very early on and the first modern cartographers represented real world objects or administrative units by accurately drawn point and line symbols that were chosen to illustrate their most important attributes. The attributes of areas were indicated by uniform colours or shading, though gradually varying shading was sometimes used to indicated the change in certain properties of the surface, such as the steepness of slopes, the depth of a lake or the variation in land cover from forest to marsh. The printing technology of etching on copper plates, available from the seventeenth century onwards, reinforced the use of a geographical symbolism that relies on well-defined, crisp delineation. Today, much geographical information concerns the location of, and interactions between, well-defined objects in space like trees in a forest, houses on a street, aeroplanes *en route* to destinations, or administrative units like the Dépendances of France (Figure 1.1*a*).

As the scientific study of the earth advanced, different kinds of attributes needed to be mapped. The study of the earth and its natural resources—geophysical geodesy, geology, geomorphology, soil science, ecology, and land—that began in the nineteenth century has continued to this day. Whereas topographical maps may be regarded as general purpose because they do not set out to fulfil any specific aim (i.e. they may be interpreted for many different purposes), maps of the distribution of rock types, soil series, or land use are made for more limited purposes. These specific-purpose maps are often referred to as 'thematic' maps because they contain information about a single subject or theme. Although these phenomena vary continuously from place to place, the surveyors were unable to record and process the huge amounts of data that are needed to provide a proper insight of their variation. Through qualitative descriptions of the variation of rocks and soil and cross-sectional diagrams the surveyors reduced their findings to sets of 'representative' taxonomic classes. Groups of related point-based taxonomic classes are linked to areas of land, called *mapping units*, which are drawn on a map using different colours or shading. Maps showing a set of delineated areas having different colours are known as *chorochromatic maps*: they are one form of the *choropleth map* or a map showing areas deemed to be of equal value (Figure 1.1*c*). To make the thematic data easy to understand choropleth maps are commonly drawn over a simplified topographic base. The use of this type of map to represent continuously varying but complex phenomena has artificially forced people to divide the space over which these phenomena occur into sets of 'objects' that are formally equivalent to the crisp entities that model well understood objects such as trees, houses, or fenced paddocks.

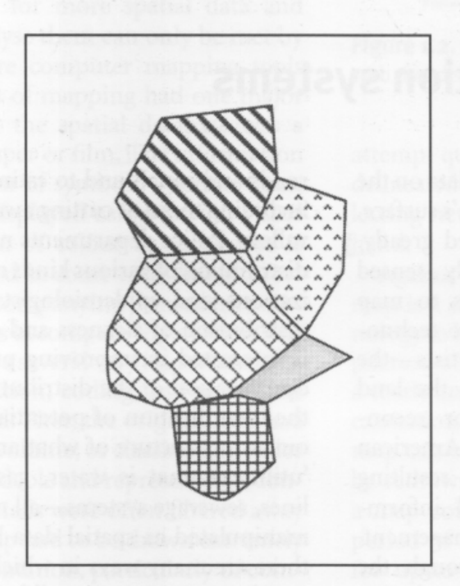
The term 'thematic map' is very widely and loosely applied (see for example, Fisher 1978, Hodgkiss 1981) and is used not only for maps showing a general purpose theme such as 'soil' or 'geology', but for maps of the distribution of a single attribute. The value of the single attribute can be linked directly to the distribution of spatial entities that are already on the map, such as the display of statistical information on population of administrative units at country, region, or local authority level.



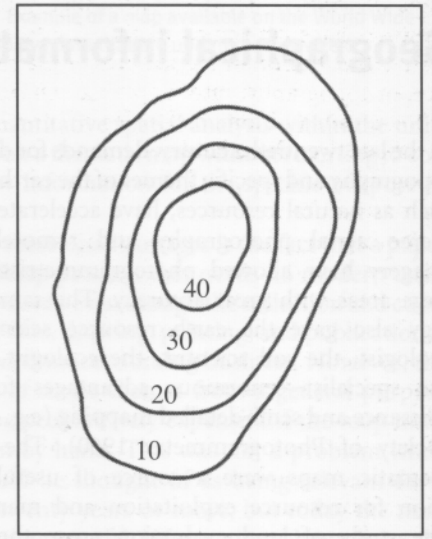
a) Objects in space



b) Continuous variation over space



c) Choropleth map



d) Isoline map

Figure 1.1. Conceptual models and representations of spatial phenomena

Certain attributes, such as the value of elevation above sea level, the soil pH over an experimental field, the variation of the incidence of a given disease in a city, or the variation of air pressure shown on a meteorological chart are more usefully represented by a continuous quantitative surface that may be modelled mathematically. The variations are conventionally shown by isolines or contours—that is lines connecting points of equal value—or by sets of monotonically increasing grey or colour scales (Figures 1.1*b*,*d*).

The first maps of the spatial distribution of rocks or soil, of plant communities or people, were qualitative. As in many new sciences, the first aim of many surveys was *inventory*—observing, classifying, and recording what is there. Qualitative methods of classification and mapping were unavoidable given the huge quantities of complex data that most environmental surveys generate. Quantitative description was not only hindered by data volume, but also by the lack of quantitative observation. Further, there was a lack of appropriate mathematical tools for describing spatial variation quantitatively. The first developments in appropriate mathematics for spatial problems came in the 1930s and 1940s in parallel with developments in statistical methods and time series analysis. Effective practical progress was completely blocked, however, by the lack of suitable computing tools. It is only since the 1960s with the availability of the digital computer that both the conceptual methods for spatial analysis and the actual potential for quantitative thematic mapping and spatial analysis have been able to blossom (e.g. Cliff and Ord 1981, Cressie 1991, Davis 1986, Journel and Huijbregts 1978, Isaaks and Srivastava 1989, Norbeck and Rystedt 1972). Today, thanks to the development of *Geographical Information Systems* (GIS) the application of logical and numerical modelling and statistical methods to spatial data is almost routine, and of great relevance, as this book explains.

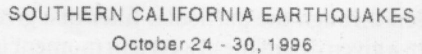
Geographical information systems

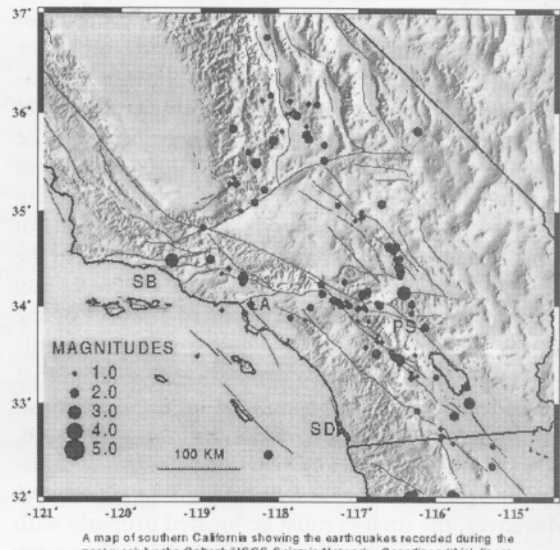
In the late twentieth century, demands for data on the topography and specific themes of the earth's surface, such as natural resources, have accelerated greatly. Stereo aerial photography and remotely sensed imagery have allowed photogrammetrists to map large areas with great accuracy. The same technology also gave the earth resource scientists—the geologist, the soil scientist, the ecologist, the land use specialist—enormous advantages for reconnaissance and semi-detailed mapping (e.g. American Society of Photogrammetry 1960). The resulting thematic maps were a source of useful information for resource exploitation and management. The study of land evaluation arose through the need to match the land requirements for producing food and supporting populations to the available resources of climate, soil, water, and technology (e.g. Brinkman and Smyth 1973, Beek 1978, FAO 1976, Rossiter 1996).

The need for spatial data and spatial analyses is not just the preserve of earth scientists. Urban planners and cadastral agencies need detailed information about the distribution of land and resources in towns and cities. Civil engineers need to plan the routes of roads and canals and to estimate construction costs, including those of cutting away hillsides and filling in valleys. Police departments need to know the spatial distribution of various kinds of crime, medical organizations and epidemiologists are interested in the distribution of sickness and disease, and commerce is interested in improving profitability through the optimization of the distribution of sales outlets and the identification of potential markets. The enormous infrastructure of what are collectively known as 'utilities'—that is water, gas, electricity, telephone lines, sewerage systems—all need to be recorded and manipulated as spatial data linked to maps. Today, there are many ways in which the political and economic processes responsible for our well-being need to be assisted such that democratic organization of the planet's resources can be used to the advantage of all. The needs are first to know what resources we have at our disposal—how much fertile land, how much energy, how many people? The second is to understand how these are distributed over the earth, and to know where they are, their extent, who owns them or has other rights to them, and how we can best use and manage them.

Now, it will always be tempting to speak of GIS and of information technology as a panacea for all urban, state, national and even global conflicts, or to think of GIS as an end rather than as a means. But we cannot become a cheerleader for a science that is disembodied from human values. The accumulation of data is meaningless unless it is underlain by a clear definition of our goals and our definitions about how we shall use and structure that science towards an informed, decision-making process.

(From a speech made by Bruce Babbit, Secretary of the United States Department of the Interior, Chairman of the Federal Data Committee, to the ESRI User Conference in Palm Springs, 21 May 1996.)





A map of southern California showing the earthquakes recorded during the past week by the Caltech/USGS Seismic Network. Coastlines (thick lines) and major faults (red lines) are marked, as well as the metropolitan areas of Los Angeles (LA), Palm Springs (PS), San Diego (SD), and Santa Barbara (SB). The circles denote the earthquakes, the size of the circle indicates the magnitude.

Figure 1.2. Example of a map available on the World Wide Web that displays data that are updated every week

The growing demands for more spatial data and for better means to analyse them can only be met by using computers. Before computer mapping tools were available, all kinds of mapping had one major limitation, namely that the spatial database was a drawing on a piece of paper or film. The information was encoded in the form of symbols—points, lines, or areas—which were displayed using various visual artifices. The symbolism of colour or text codes is explained in a legend and in some cases more information is given in an accompanying printed memoir.

The paper map and its accompanying memoir was *the* database so there were several very important consequences for the collection, coding, and use of the information it contained. First, the original data had to be greatly reduced in volume, or classified, in order to make them understandable and representable; consequently many local details were often filtered away and lost. Second, the map had to be drawn extremely accurately and the presentation, particularly of complex themes, had to be very clear. Third, the sheer volume of information meant that areas that are large with respect to the map scale could only be represented by a number of map sheets. It is a common experience that one's area of interest is frequently near the junction of two, if not more, map sheets! Fourth, once data had been put into a map, it was not cheap nor easy to retrieve them in order to combine them with other spatial data. Fifth, the printed map is a static, qualitative document. It is extremely difficult to attempt quantitative spatial analysis within the units delineated on a thematic map without resorting to collecting new information for the specific purpose in hand.

Against these disadvantages is the fact that a paper map is a cheap product that needs no modern technology to be read even though the collection and compilation of data, drawing, printing, and publication is a costly and time-consuming business. However, the extraction of single themes from a general purpose map can be prohibitively expensive if the map must be redrawn by hand. This was not a problem when a map could be thought of as being relevant for a period of twenty years or more, but today the need for up-to-date spatial information about how the earth's surface is changing means that conventional map-making techniques are totally inadequate. For example, for some kinds of mapping such as weather charts, information on earthquakes, or the distribution net of a telephone company, there may be a daily, or even hourly need for the spatial database to be brought up to date, which is simply not possible by hand (e.g. Figure 1.2, Plates 2.2, 4.1).

Essentially, the hand-drawn map or the map in a resource inventory is merely a snapshot of the situa-

tion seen through the particular filter of a given surveyor in a given discipline at a certain moment in time. More recently, the aerial photograph, but more especially the satellite image, have made it possible to see how landscapes change over time, to follow the relentless march of desertification or erosion or the swifter progress of forest fires, floods, locust swarms, or weather systems. But the products of the airborne and space sensors are not maps, in the original meaning of the word, but photographic images or streams of data on magnetic tapes. The digital data are not in the familiar form of points, lines, and areas representing the already recognized and classified features of the earth's surface, but are coded in picture elements —pixels—cells in a two-dimensional matrix that contain merely a number indicating the strength of reflected electromagnetic radiation in a given band. New tools were needed to turn these streams of numbers into pictures and to identify meaningful patterns. Moreover, the images must be located properly with respect to a geodetic grid, otherwise the data they carry cannot be related to a definite place. The need arose for a marriage between remote sensing, earth-bound survey, and cartography; this has been made possible by the class of spatial information handling and mapping tools known as geographical information systems.

Alternatives for Handling Complex Data: Generalization or Analysis

During the 1960s and 1970s there were new trends in the ways in which data about natural resources of soil and landscape systems were being used for resource assessment, land evaluation, and planning. Realizing that different aspects of the earth's surface do not function independently from each other, and at the time having insufficient means to deal with huge amounts of disparate data, people attempted to evaluate them in an integrated, multidisciplinary way. The 'gestalt' method (Hills 1961, Hopkins 1977, Vink 1981) classifies the land surface into what are presumed to be 'naturally occurring' environmental units or building blocks that can be recognized, described, and mapped in terms of the total interaction of the attributes under study. Within these 'natural units' there is supposed to be a recognizable, unique, and interdependent combination of the environmental characteristics of landform, geology, soil, vegetation, and water. The same basic idea was used in integrated resource surveys—classic examples come from the Australia integrated resource or land systems surveys (Christian and Stewart 1968) or from the former UK Land Resources Division (e.g. Brunt 1967) and work at the ITC in Enschede in the Netherlands. This method addresses the problem of excessive amounts of data by reducing all spatial variation to a limited number of supposedly homogeneous classes that can be drawn on a choropleth map.

The integrated survey approach has several inherent problems and it has largely fallen into disuse. In many cases the information on the supposedly homogeneous units was too general for many applications and it was very difficult or impossible to retrieve specific information about particular aspects of a landscape. The division of the landscape into spatial units also depends greatly on personal insight, and the level of resolution of the survey had more to do with the scale of the topographical base map used than with the level of survey needed to map specific attributes. So a ready market remains for the more conventional monodisciplinary surveys, such as those of geology, landform, soil, vegetation, and land use. Increasing pressure on land has inflated the demand for location-specific information so surveys have tended towards collecting more and more data. The concomitant increase in data volumes and new insights for analysing spatial patterns and processes has in its turn stepped up the search for simple, reproducible methods for combining data from several different sources and levels of resolution.

Early on, planners and landscape architects, particularly in the United States of America, realized that data from several monodisciplinary resource surveys could be combined and integrated simply by overlaying transparent copies of the resource maps on a light table, and looking for the places where the boundaries on the several maps coincided. One of the best-known exponents of this simple technique was the American landscape architect Ian McHarg (McHarg 1969). In 1963, another American architect and city planner, Howard T. Fisher, elaborated Edgar M. Horwood's idea of using the computer to make simple maps by printing statistical values on a grid of plain paper (Sheehan 1979). Fisher's program SYMAP, short for SYnagraphic MAPping system (the name has its origin in the Greek word *synagein*, meaning to bring together), includes a set of modules for analysing data, and manipulating them to produce choropleth or isoline interpolations, with the results to be displayed in many ways using the overprinting of line-printer characters to produce suitable grey scales.

Fisher became director of the Harvard Graduate School of Design's Laboratory for Computer

BOX 1.1.

Arguments for computer cartography

1. To make existing maps more quickly
2. To make existing maps more cheaply
3. To make maps for specific user needs
4. To make map production possible in situations where skilled staff are unavailable
5. To allow experimentation with different graphical representations of the same data
6. To facilitate map making and updating when the data are already in digital form
7. To facilitate analyses of data that demand interaction between statistical analyses and mapping
8. To minimize the use of the printed map as a data store and thereby to minimize the effects of classification and generalization on the quality of the data
9. To create maps that are difficult to make by hand, e.g. 3D maps or stereoscopic maps
10. To create maps in which selection and generalization procedures are explicitly defined and consistently executed
11. Introduction of automation can lead to a review of the whole map-making process, which may also lead to savings and improvements

Graphics, and SYMAP was first in a line of mapping programs that were produced by an enthusiastic, internationally well-known, and able staff. Among these programs were the grid cell (or raster) mapping programs GRID and IMGRID that allowed the user to do in the computer what McHarg had done with transparent overlays. Naturally, the Harvard group was not alone in this new field and many other workers developed programs with similar capabilities (e.g. Duffield and Coppock 1975, Steiner and Matt 1972, Fabos and Caswell 1977 to name but a few). Initially none of these overlay programs allowed the user to do anything that McHarg could not do; they merely speeded up the process and made it reproducible. However, users soon began to realize that with little extra programming effort, they could do other kinds of spatial and logical analysis on mapped data that were helpful for planning studies (e.g. Steinitz and Brown 1981) or ecological analysis (e.g. Luder 1980); previously these computations had been extremely difficult to do by hand. We note in passing, that although the terminology and the sources of the data are quite different, many of the image analysis methods used in these raster map analysis programs are little different from those first adopted for processing remotely sensed data.

Because SYMAP, GRID, IMGRID, GEOMAP, MAP, and many of these other relatively simple programs were designed for quick and cheap analysis of gridded data and their results could only be displayed by using crude lineprinter graphics, many cartographers refused to accept the results they produced as maps. Cartographers had begun to adopt computer techniques in the 1960s, but until recently they were largely limited to aids for the automated drafting and preparation of masters for printed maps. For traditional cartography the new computer technology did not change fundamental attitudes to map-making—the high-quality paper map remained both the principal data store and the end-product. However, by 1977 the experience of using computers in map making had advanced so far that Rhind (1977) was able to present many cogent reasons for using computers in cartography (Box 1.1).

Investments in computer cartography

By the late 1970s there had been considerable investments in the development and application of computer-assisted cartography, particularly in North America by government and private agencies (e.g. Tomlinson *et al.* 1976, Teicholz and Berry 1983). In Europe and Australasia the developments proceeded on a smaller scale than in North America but during the 1970s and early 1980s major strides in using and developing computer-assisted cartography were made by several nations, notably Sweden, Norway, Denmark, France, the Netherlands, the United Kingdom, West Germany, and Australia. Literally hundreds of computer programs and systems were developed for

various mapping applications, mostly in government institutes and universities.

The introduction of computer-assisted cartography did not immediately lead to a direct saving in costs as had been hoped. The acquisition and development of the new tools was often very expensive; computer hardware was extremely expensive; there was a shortage of trained staff and many organizations were reluctant or unable to introduce new working practices. Initially, the computer-assisted mapping market was seen by many manufacturers of computer-aided design and computer graphics systems as so diverse that the major investments needed to develop the software were unlikely to reap returns from a mass market. Consequently, many purchasers of expensive systems were forced to hire programming staff to adapt a particular system to their needs.

The history of using computers for mapping and spatial analysis shows that there have been parallel developments in automated data capture, data analysis, and presentation in several broadly related fields such as cadastral and topographical mapping, thematic cartography, civil engineering, geology, geography, hydrology, spatial statistics, soil science, surveying and photogrammetry, rural and urban planning, utility networks, and remote sensing and image analysis. Military applications have overlapped and even dominated several of these monodisciplinary fields. Consequently there has been much duplication of effort and a multiplication of discipline-specific jargon for different applications in different lands. This multiplicity of effort in several initially separate, but closely related fields has resulted in the emergence of the general purpose GIS.

During the 1990s there were several important technical and organizational developments that greatly assisted the wide application and appreciation of GIS. The first is awareness—many more people now know why it is important to be able to manipulate large amounts of spatial information efficiently, though many more still need to be convinced. Much knowledge has been built up on how to set up computer mapping and GIS projects efficiently (Huxhold and Levinsohn 1995). Second, by 1995 computer technology had provided huge amounts of processing power and data storage capacity on modestly priced personal computers enabling GIS to be used by individuals and organizations with limited budgets. Third, many computers are connected by electronic networks, allowing expensive data and software to be shared. Fourth, standardization in interfaces between database programs and other computer programs has made it much easier to provide the functionality for handling large amounts of data easily. Fifth, the basic functionality required for handling spatial data has been widely accepted to the point where a limited number of commercial systems dominate the marketplace, thereby bringing a major degree of uniformity to a previously heterogeneous domain. This last point is particularly true for the automation of existing techniques. However, in spite of the impressive technical developments, the impacts of GIS on the development of a fundamental body of theory of spatial interactions have been limited. As will be shown in Chapter 2 and elsewhere in this book, the basic spatial models used in modern GIS are little different from those used 15–20 years ago; new insights into spatial and temporal modelling still need to be made if GIS is to develop beyond a mere technology.

'strengthening a particular technique [i.e. introducing automation]—putting muscles on it—contributes nothing to its validity. The poverty of the technique, if it is indeed impotent to deal with its presumed subject-matter, is hidden behind a mountain of effort . . . the harder the subproblems were to solve, and the more technical success was gained in solving them, the more is the original technique fortified'. (Weizenbaum 1976)

The main result of more than twenty years of technical development is that GIS have become a worldwide phenomenon. In 1995 it was estimated that these technical solutions had been installed at more than 93 000 sites worldwide and while North America and Europe dominate this user base (65 per cent and 22 per cent respectively, Estes 1995) many other countries are now beginning to exploit GIS capabilities. Today (1997), these systems are used in many different fields (Box 1.2); since 1986 there have been some 200 books written on various aspects of the subject, there have been literally hundreds of conferences, and there are several important scientific and trade journals devoted entirely to the design, technology, use, and management of GIS. There have been major pronouncements by international and national governments about the importance of spatial information in modern society (e.g. Department of Environment 1987) for planning, marketing, and the development of the 'information society'.

BOX 1.2. ACTIVE DOMAINS FOR GIS

Producers, kinds, and applications of Geographical Information

The main producers and sources

Topographical Mapping: National Mapping Agencies, private Mapping Companies
Land Registration and Cadastre
Hydrographic Mapping
Military Organizations
Remote Sensing companies and satellite agencies
Natural resource surveys: Geologists; Hydrologists; Physical Geographers and Soil Scientists; Land Evaluators; Ecologists and Biogeographers; Meteorologists and Climatologists; Oceanographers

The main types of geographical data available

Topographic maps at a wide range of scales
Satellite and airborne scanner images and photographs
Administrative boundaries: Census tracts and census data; Postcode areas
Statistical data on people, land cover, land use at a wide range of levels.
Data from marketing surveys.
Data on utilities (gas, water, electricity lines, cables) and their locations.
Data on rocks, water, soil, atmosphere, biological activity, natural hazards, and disasters collected for a wide range of spatial and temporal levels of resolution.

Some current applications

Agriculture	Monitoring and management from farm to National levels
Archaeology	Site description and scenario evaluation
Environment	Monitoring, modelling, and management for land degradation; land evaluation and rural planning; landslides; desertification; water quality and quantity; plagues; air quality; weather and climate modelling and prediction.
Epidemiology and Health	Location of disease in relation to environmental factors
Forestry	Management, planning, and optimizing extraction and replanting
Emergency services	Optimizing fire, police, and ambulance routing; improved understanding of crime and its location.
Navigation	Air, sea, and land.
Marketing	Site location and target groups; optimizing goods delivery
Real Estate	Legal aspects of the cadastre, property values in relation to location, insurance
Regional/local planning	Development of plans, costing, maintenance, management.
Road and rail	Planning and management
Site evaluation and costing	Cut and fill, computing volumes of materials
Social studies	Analysis of demographic movements and developments
Tourism	Location and management of facilities and attractions
Utilities	Location, management, and planning of water, drains, gas, electricity, telephone, cable services

The structure of this book

It is not the aim of this book to describe the growth of awareness of the advantages of GIS among politicians, businessmen, academics, and resource managers, nor to explain how these tools can be efficiently incorporated in modern government or business. Others have performed this task (e.g. Huxhold and Levinsohn 1995) and the job is ongoing. As with the first edition (Burrough 1986), the aim of this book is to explain the scientific and technical aspects of working with geographical information, so that users of GIS understand the general principles, opportunities, and pitfalls of recording, collecting, storing, retrieving, analysing, and presenting spatial information. The aim includes not only the basic principles but also an understanding of the limitations of the technology and the role of errors in data collection, conceptualization, analysis, and presentation on the perception of the results. An old computer adage of *garbage in, garbage out!* holds true in GIS use. The very complexity of these systems makes it possible to input good data but deliberately or unwittingly to produce garbage which, thanks to multi-colour high-resolution displays looks like a high-quality product. At the same time, it is not difficult to conceal bad data through clever or high-quality presentation. The problem is not the technology, but one of the complexity of spatial information (Monmonnier 1993); with computers one can make bigger and better mistakes faster than ever.

This book provides a comprehensive, but of necessity limited, overview of the main aspects of handling spatial information in a GIS. In this chapter we present the history of development and the basic structure of GIS. In Chapter 2 we explore the problems of collecting and describing spatial data and the dilemma of dealing with discrete entities in space (the reductionist paradigm) or with the continuous variation of an attribute of space. How should we proceed when we are unsure which model to use? Which approaches are preferred by which disciplines, and how can the observational and conceptual models be combined? In Chapter 3 we explore ways in which the different paradigms used in Chapter 2 can be implemented technically in the computer, for storage, retrieval, and display.

The practical aspects of building a geographical database of entities in space are described in Chapter 4, which covers the essential aspects of digitizing, scanning, storing, updating, and plotting geographical data according to the entity paradigm, which is probably the most common, particularly in applications involving topographic and cadastral mapping. The creation of databases of attributes modelled by continuous fields is described in Chapters 5 and 6, which also deal with the input and rectification of remotely sensed images, the derivation of digital elevation models from aerial photographs and other data sources where cover is reasonably complete, and the interpolation of data from point observations of quantitative and qualitative attributes to continuous fields.

Data analysis and numerical modelling are explored in Chapters 7 and 8. Following the two paradigms of spatial information, Chapter 7 concentrates on the analysis of entity-based data—numerical and statistical operations on attributes, proximity relations, and network interactions. Chapter 8 explores the many ways in which new information can be derived from continuous fields. Both chapters present practical examples to illustrate the theory.

Many information systems present a false sense of security. The high-quality graphics presentation and exactly formulated query languages often insulate the user from real questions about the quality of the data, the levels of errors that occur, and the levels of confidence that should be associated with the results. Chapters 9 and 10 explore the ways in which errors can occur in the data, the effects they may have on the results of analyses, and the ways in which statistical treatment of spatial errors can be used to improve the information content rather than reduce it.

Up to this point all discussions will have been in terms of one spatial paradigm or the other, and the ways they may be combined (as for example, by mapping the variations of annual rainfall (modelled as a continuous field) over a continent (modelled as a set of country entities with crisp boundaries). The choice of paradigm is made difficult because different scientists and practitioners often tend to describe the same area in different ways. Even if they stem from the same discipline they tend to draw boundaries around geographical 'objects' in different ways and these boundaries are usually convoluted and highly irregular (Legros *et al.* 1996). The problem is worse when data from surveys made at different levels of resolution have to be combined. Consequently geographical phenomena, at least as described by current generations of natural and social scientists, cannot be described uniquely.

New theoretical and practical research on fuzzy logic and continuous classification suggests that it is not always necessary to force spatial data into either one or the other paradigm—the real world is not made up of either crisp entities or continuous fields—but many phenomena have properties that can be dealt with by combining the best aspects of either. Previously it was thought impossible to deal with vague formulations of space in the computer, but Chapter 11 shows that this is far from being so. In fact, by refusing to be hidebound by the classical paradigms we find that we can often make more specific pronouncements about spatial relations than we can with conventional methods. This finding provides many new opportunities for useful research and applications that will take many more years to work out fully.

Definitions of GIS

The tool-base definition of a GIS is *a powerful set of tools for collecting, storing, retrieving at will, transforming and displaying spatial data from the real world for a particular set of purposes.* The geographical (or spatial) data represent phenomena from the real world in terms of (*a*) their position with respect to a known coordinate system, (*b*) their attributes that are unrelated to position (such as colour, cost, pH, incidence



BOX 1.3. DEFINITIONS OF GIS

Definitions of GIS

(*a*) Toolbox-based definitions.

'*a powerful set of tools for collecting, storing, retrieving at will, transforming and displaying spatial data from the real world*' (Burrough 1986).

'*a system for capturing, storing, checking, manipulating, analysing and displaying data which are spatially referenced to the Earth*' (Department of Environment 1987).

'*an information technology which stores, analyses, and displays both spatial and non-spatial data*' (Parker 1988).

(*b*) Database definitions

'*a database system in which most of the data are spatially indexed, and upon which a set of procedures operated in order to answer queries about spatial entities in the database*' (Smith *et al.* 1987).

'*any manual or computer based set of procedures used to store and manipulate geographically referenced data*' (Aronoff 1989).

(*c*) Organization-based definitions.

'*an automated set of functions that provides professionals with advanced capabilities for the storage, retrieval, manipulation and display of geographically located data*' (Ozemoy, Smith, and Sicherman 1981).

'*an institutional entity, reflecting an organisational structure that integrates technology with a database, expertise and continuing financial support over time*' (Carter 1989).

'*a decision support system involving the integration of spatially referenced data in a problem solving environment*' (Cowen 1988).

of disease, etc.) and (*c*) their spatial interrelations with each other which describe how they are linked together (this is known as topology and describes space and spatial properties such as connectivity which are unaffected by continuous distortions).

Others have provided alternative definitions of GIS (Box 1.3), focusing on either the spatial *database* or on the *organizational* aspects. The database definition emphasizes the differences in data organization needed to handle spatial data with their location, attributes, and topology and most other kinds of information that only have to do with entities and attributes. The organizational definition emphasizes the role of institutes and people in handling spatial information rather than the tools they need. In this sense GIS have been serving society for a very long time.

The components of a geographical information system

Geographical information systems have three important components—computer hardware, sets of application software modules, and a proper organizational context including skilled people—which need to be in balance if the system is to function satisfactorily.

Computer hardware

The general hardware components of a geographical information system are presented in Figure 1.3. The computer has a hard disk drive for storing data and programs, but extra storage can be provided via a network or by digital tape cassettes, optical CD-ROMs, and other devices. A digitizer or a scanner is used to convert maps and documents into digital form so that they can be used by the computer programs. A plotter or a printer or any other kind of display device is used to present the results of the data processing. Inter-computer communication is provided by local and global electronic networks using special data lines with optical fibres or over ordinary telephone lines by using a device known as a 'modem'. The user controls the computer and the peripherals (a general term for plotters, printers, digitizers, and other apparatus

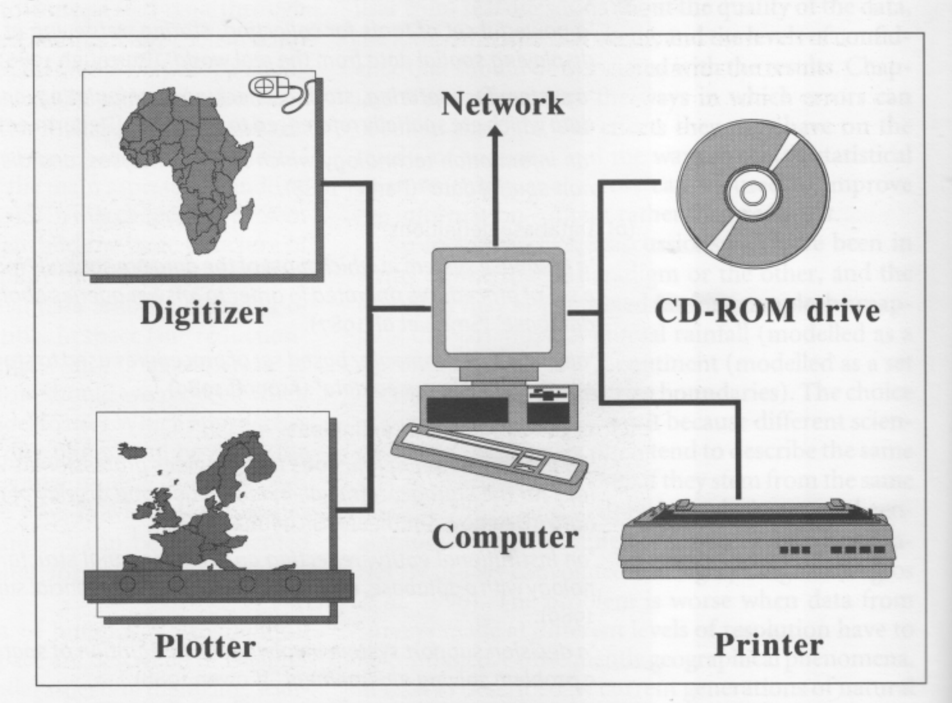


Figure 1.3. The major hardware components of a geographical information system

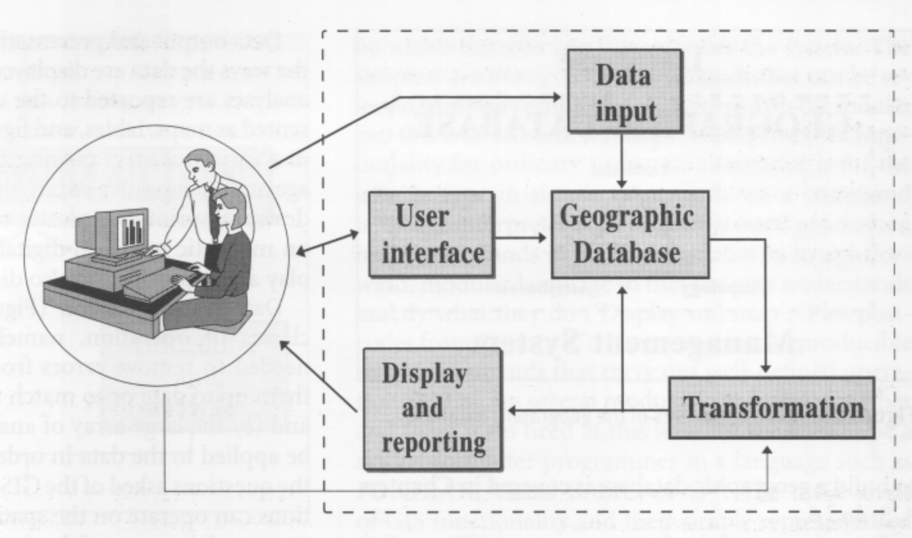


Figure 1.4. The main software components of a geographical information system

linked to the computer) via the computer screen and keyboard, aided by a 'mouse' or pointing device.

GIS SOFTWARE

The software for a geographical information system may be split into five functional groups (Figure 1.4):

(*a*) Data input and verification
(*b*) Data storage and database management
(*c*) Data output and presentation
(*d*) Data transformation
(*e*) Interaction with the user.

Data input (Figure 1.5) covers all aspects of capturing spatial data from existing maps, field observations, and sensors (including aerial photography, satellites, and recording instruments) and converting them to a standard digital form. Many tools are available including the interactive computer screen and mouse, the digitizer, word processors and spreadsheet programs, scanners (in satellites or aeroplanes for direct recording of data or for converting maps and photographic images), and devices necessary for reading data already written on magnetic media such as tapes or CD-ROMs. Data input, and the verification of data needed

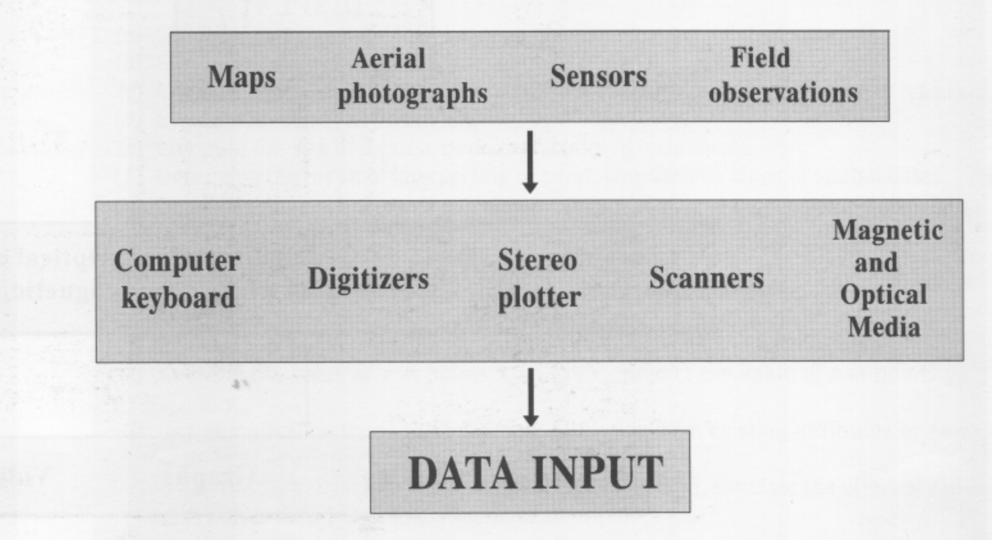


Figure 1.5. Data collection and input

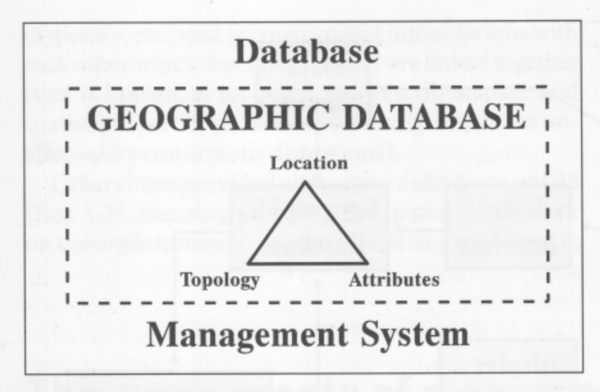


Figure 1.6. The components of the geographic database

to build a geographic database is covered in Chapters 4, 5, and 6.

Data storage and database management (Figure 1.6) concerns the way in which data about the location, linkages (topology), and attributes of geographical elements (points, lines, areas, and more complex entities representing objects on the earth's surface) are structured and organized, both with respect to the way they must be handled in the computer and how they are perceived by the users of the system. The computer program used to organize the database is known as a Database Management System (DBMS). Data models (formalized descriptions of real world phenomena), database structures, and methods of database organization are discussed in Chapters 2 and 3.

Data output and presentation (Figure 1.7) concern the ways the data are displayed and how the results of analyses are reported to the users. Data may be presented as maps, tables, and figures (graphs and charts) in a variety of ways ranging from the ephemeral image on the computer screen, through hardcopy output drawn on printer or plotter to information recorded on magnetic media in digital form. Methods of display and reporting are also discussed in Chapter 4.

Data transformation (Figure 1.8) embraces two classes of operation, namely (*a*) transformations needed to remove errors from the data or to bring them up to date or to match them to other data sets, and (*b*) the large array of analysis methods that may be applied to the data in order to achieve answers to the questions asked of the GIS (Box 1.4). Transformations can operate on the spatial, topological, and the non-spatial aspects of the data, either separately or in combination. Many of these transformations, such as those associated with scale changing, fitting data to new projections, logical retrieval of data, and calculation of areas and perimeters are of such a general nature that one should expect to find them in every kind of GIS in one form or another. Other kinds of manipulation may be extremely application-specific, and their incorporation into a particular GIS may be only to satisfy the particular users of that system. The kinds of transformation methods available, their optimum use and misuse, the ways in which sets of simple transformations may be combined in order

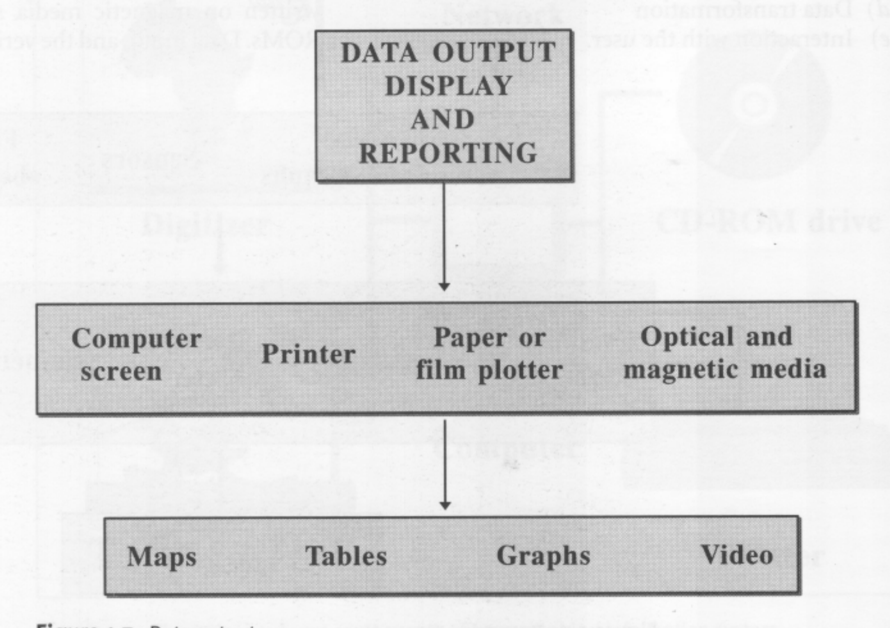


Figure 1.7. Data output

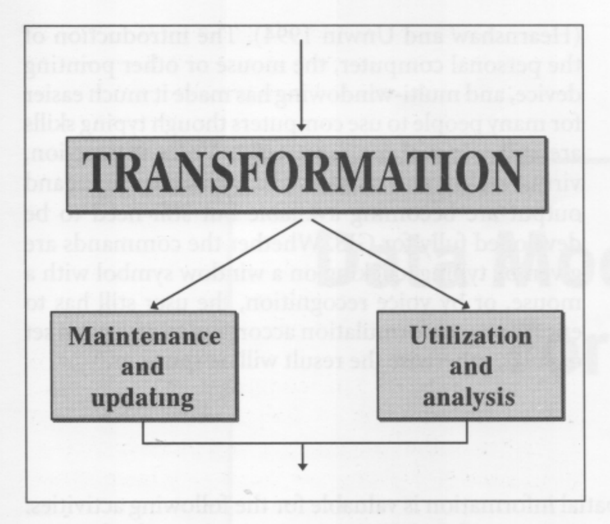


Figure 1.8. Data transformation

to achieve certain types of geographical or spatial modelling are the major subject of this book, being covered by Chapters 7, 8, and 11; certain spatial transformations and transformations necessary to ensure the integrity of a database are discussed in Chapter 4.

GIS designers have realized that the requirements of users to retrieve and transform data are unlimited. Therefore most systems provide a range of interfaces by which the user can interact with the system. The simplest are menu-driven commands that can be selected by simply pointing and clicking with the mouse, and this is an efficient way of providing complex functionality for ordinary users. An alternative is for the user to type in simple commands via a command language interpreter (CLI). As with word processors, many commands in a GIS are written in straightforward, modular language so that the user understands exactly what they do ('Display soil map'; 'Plot postcodes from County X'). Users can create reproducible sets of commands that carry out well-defined operations by linking several modular commands together and there is no need at this level for the user to be a skilled computer programmer in a language such as FORTRAN, Visual BASIC, or C^{++}. The main kinds of GIS functionality and their simple representation are explained in Chapters 5, 6, 7, and 8, together with examples of their use.

Of course, not all operations can be carried out using the basic menu-driven commands, and users may have to write their own computer programs to meet their needs. Some GIS systems provide so-called Macro Languages—simplified, formal programming languages that may be used to link many basic applications together. The latest developments, particularly for simulation modelling with GIS, are to provide

BOX 1.4.

Basic requirements for a GIS

(*a*) Show the locations of entities of type A.
(*b*) Show the location of entity A in relation to place B.
(*c*) Count the number of occurrences of entity type A within distance D of entity type B.
(*d*) Evaluate function *f* at position X.
(*e*) Compute the size of B (area, perimeter, count of inclusions).
(*f*) Determine the result of intersecting or overlaying various kinds of spatial data.
(*g*) Determine the path of least cost, resistance or distance along the ground from X to Y over a network or a continuous surface.
(*h*) List the attributes of entities located at points X1, X2.
(*i*) Determine which entities are next to entities having certain combinations of attributes
(*j*) Reclassify or recolour entities having certain combinations of attributes.
(*k*) Knowing the value of z at points x_1 x_2, . . . x_n, predict the value of z at points y_1, y_2, . . . y_m.
(*l*) Use numerical methods to derive new attributes from existing attributes or new entities from existing entities.
(*m*) Using the digital database as a model of the real world, simulate the effect of process P over time T for a given scenario S.

the user with a high-level, compilable programming language in which the models can be efficiently written (cf. PCRaster—Wesseling *et al.* 1996, or MMS—Leavesley *et al.* 1996). In other situations the GIS is used to assemble the data for a complex model that is programmed outside the system using a standard computer language (Burrough 1996*a*). Once the model has been run, the results are transferred back to the GIS for display.

The interaction between user and GIS for data and query input and the writing of models for data analysis is an aspect that has been neglected until recently (Hearnshaw and Unwin 1994). The introduction of the personal computer, the mouse or other pointing device, and multi-windowing has made it much easier for many people to use computers though typing skills are still essential for most tasks. Voice interaction, virtual reality, and multi-media with sound input and output are becoming available but still need to be developed fully for GIS. Whether the commands are given by typing, clicking on a window symbol with a mouse, or by voice recognition, the user still has to ensure proper formulation according to an agreed set of rules, otherwise the result will be spurious.

Questions

1. Explain why up-to-date spatial information is valuable for the following activities: marketing, real estate, banking, tourism, electricity generation and supply, forestry, agriculture.

2. Why should a national government provide spatial data on the World Wide Web for free? Examine the aims and motives behind this approach.

3. Explore the differences between digital cartography, geographical information, and computer-aided design and explain why each has a different market niche.

Suggestions for further reading

AGI (1996). *AGI online Glossary of GIS.* The Internet. *(http://www.geo.ed.ac.uk/root/agidict/html/welcome.html)*

Goodchild, M. F., Steyaert, L. T., Parks, B. O., Johnston, C., Maidment, D., Crane, M., and Glendinning, S. (eds.) (1996). *GIS and Environmental Modeling: Progress and Research Issues.* GIS World Books, Fort Collins, Colo., 486 pp.

Huxhold, W. E. (1991). *An Introduction to Urban Geographic Information Systems.* Oxford University Press, New York, 337 pp.

—— and Levinsohn, A. G. (1995). *Managing Geographic Information System Projects.* Oxford University Press, New York, 247 pp.

McDonnell, R. A., and Kemp, K. K. (1995). *The International GIS Dictionary.* GeoInformation International, Cambridge.

Machover, C. (1989). *The C4 Handbook: CAD, CAM, CAE, CIM.* Tab Books Inc., Blue Ridge Summit, Pa., 438 pp.

Maguire, D. J., Goodchild, M. F., and Rhind, D. (eds.) (1991). *Geographical Information Systems: Principles and Applications.* Longman Scientific and Technical, Harlow, 2 vols.

Masser, I., and Blakemore, M. (eds.) (1991). *Handling Geographical Information: Methodology and Potential Applications.* Longman Scientific and Technical, Harlow, Essex, 317 pp.

Peuquet, D. J., and Marble, D. F. (eds.) (1990). *Introductory Readings in Geographic Information Systems.* Taylor & Francis, London, 371 pp.

Raper, J., Rhind, D., and Shepherd, J. (1992). *Postcodes: The New Geography.* Longman Scientific and Technical, Harlow, Essex, 322 pp.

TWO

Data Models and Axioms: Formal Abstractions of Reality

When someone views an environment they simplify the inherent complexity of it by abstracting key features to create a 'model' of the area. This cognitive exercise is influenced by the cultural norms of the observer and the purpose of the study. This chapter examines the various model development stages that take place in the process of producing geographical data that may be used by others in a graphical or digital form. It is important to examine these theoretical ideas as all the data we use in a GIS will have been schematized using these geographical data models.

The two extremes in approach perceive space either as being occupied by a series of entities which are described by their properties and mapped using a co-ordinate system, or as a continuous field of variation with no distinct boundaries. Formalized geographical data models are used to characterize these conceptual ideas so that they may be broken down into units which may be recorded and mapped. The principal approaches use either a series of points, lines, and polygons, or tessellated units to describe the various features in a landscape. The adoption of a particular model influences the type of data that may be used to describe the phenomena and the spatial analysis that may be undertaken. The fundamental procedures and axioms for handling and modifying spatial data are explained. Practical examples of the choice and use of various data models in frequently encountered applications are given.

Imagine that you are talking on the telephone to someone and they ask you to describe the view from your window. How would you depict the variations you see? It is likely that you would break down the landscape into units such as a building, road, field, valley, or hill and use geographical referencing in terms of 'beside', 'to the left of', or 'in front of' to describe the features. You have in fact developed a conceptual model of the

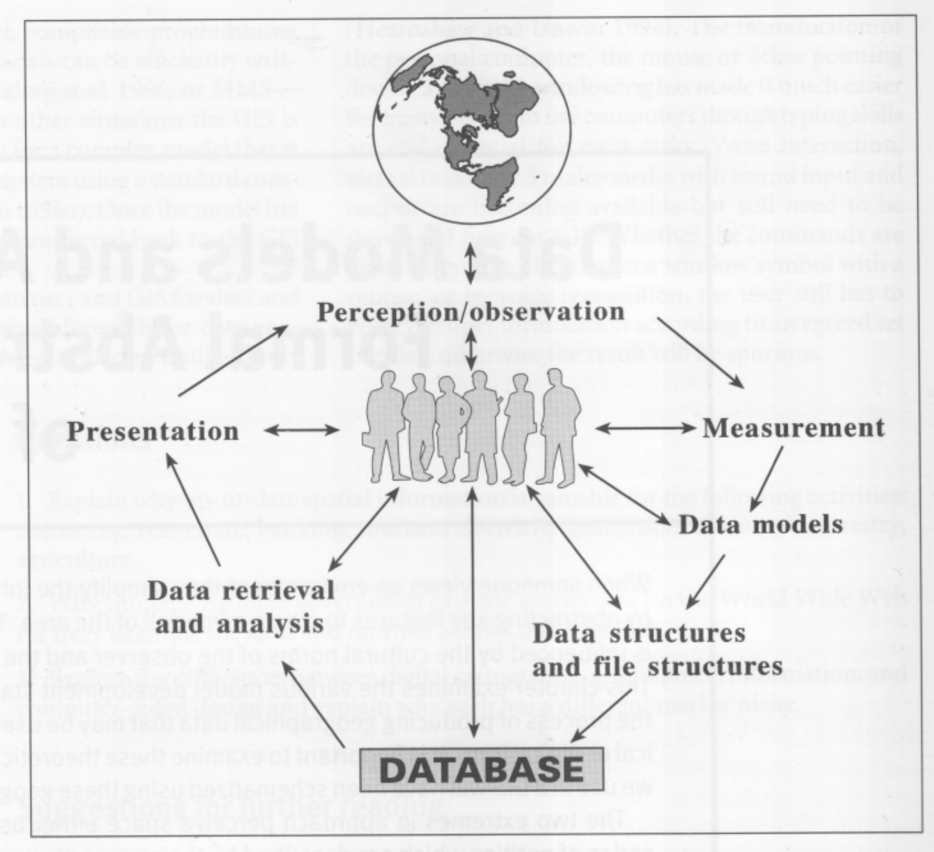


Figure 2.1. All aspects of dealing with geographical information involve interactions with people

BOX 2.1. SPATIAL DATA MODELS AND DATA STRUCTURES

Spatial data models and data structures

The creation of analogue and digital spatial data sets involves seven levels of model development and abstraction (cf. Peuquet 1984*a*, Rhind and Green 1988, Worboys 1995):

(*a*) A view of reality (conceptual model)
(*b*) Human conceptualization leading to an analogue abstraction (analogue model)
(*c*) A formalization of the analogue abstraction without any conventions or restrictions on implementation (spatial data model)
(*d*) A representation of the data model that reflects how the data are recorded in the computer (database model)
(*e*) A file structure, which is the particular representation of the data structure in the computer memory (physical computational model).
(*f*) Accepted axioms and rules for handling the data (data manipulation model)
(*g*) Accepted rules and procedures for displaying and presenting spatial data to people (graphical model)

landscape. Your interpretation of the features you have observed and the ones you have decided to ignore will be influenced by your experience, your cultural background, and that of the person to whom you are describing the scene.

When information needs to be exchanged over a larger domain it becomes necessary to formalize the models used to describe an area to ensure that data are interpreted without ambiguity and communicated effectively. This chapter will describe the main data models used for describing geographical phenomena (see Couclelis 1992, Frank *et al.* 1992; Frank and Campari 1993; Egenhofer and Herring 1995; and Burrough and Frank 1996 for more detailed discussion). It gives an essential background to the following chapters of this book, because we do not store real world phenomena in the computer but only representations based on these formalized models. The major steps involved in proceeding from human observation of the world, either directly or with the assistance of tools like aerial photographs, remotely sensed images, or statistically located samples, to an *analogue* or digital representation are outlined in Box 2.1 and illustrated in Figure 2.1. The most important first step is that people observe the world and perceive phenomena that are fixed or change in space and time. Their perception will influence all subsequent analysis; success or failure with GIS does not depend in the first instance on technology but more on the appropriateness or otherwise of the conceptual models of space and spatial interactions.

Conceptual models of real world geographical phenomena

Geographical phenomena require two descriptors to represent the real world; *what* is present, and *where* it is. For the former, phenomenological concepts such as 'town', 'river', 'floodplain', 'ecotope', 'soil association' are used as fundamental building blocks for analysing and synthesizing complex information. These phenomena are recognized and described in terms of well-established 'objects' or 'entities', which are defined in standard texts (cf. Goudie *et al.* 1988, Johnston *et al.* 1988, Lapedes 1976, Lapidus 1987, Scott 1980, Stevens 1988, Whitten and Brooks 1972, Whittow 1984). However, these dictionaries fail to point out that there are many ways to describe these phenomena, and different terms can be used for different levels of resolution. Many of these perceived geographical phenomena described by people as explicit entities (such as 'hill', 'town', or 'lake') do not have an exact form and their extent may change with time (e.g. see Burrough and Frank 1996).

At the same time, the type of building block used to describe a phenomena at one scale of resolution is likely to be quite different from that at another. For example, a road imaged from a satellite-based sensor might be modelled as a line, but the plan of a building site would have to be modelled using an areal representation to show its various structures. Phenomena are also very often grouped or divided into units at other levels of resolution ('scales') according to hierarchically defined taxonomies; for example the hierarchy of administration units of country–province–town–district, or of most soil, plant, or animal classification systems.

The referencing in space of the phenomena may be defined in terms of a geometrically exact or a relative location. The former uses local or world coordinate systems defined using a standard system of spheroids, projections, and coordinates which give an approximation of the form of the earth (a spheroid) onto a flat surface. The coordinate system may be purely local, measured in tens of metres, or it may be a national grid or an internationally accepted projection that uses geometrical coordinates of latitude and longitude. Alternatively some maps provide geographical referencing in a relative, rather than an absolute spatial geometry as illustrated by aboriginal rock paintings and the plan of the London Underground. With these maps the locations are defined in reference to other features within the space, and neighbourhoodness and direction between entities is shown rather than actual metric distances.

Conceptual models of space: entities or fields

Is the geographic world a jig-saw puzzle of polygons, or a club-sandwich of data layers? (Couclelis 1992)

From these conceptual ideas of geographical phenomena it is possible to formalize the representation of space and spatial properties. When considering any space—a room, a landscape, or a continent—we may adopt several fundamentally different ways to describe what is going on in that subset of the earth's surface. The two extremes are (*a*) to perceive the space as being occupied by *entities* which are described by their attributes or properties, and whose position can be mapped using a geometric coordinate system, or (*b*) to imagine that the variation of an attribute of interest varies over the space as some continuous mathematical function or *field*.

Entities. The most common view is that space is peopled with 'objects' (entities). Defining and recognizing the entity (is it a house, a cable, a forest, a river, a mountain?) is the first step; listing its attributes, defining its boundaries and its location is the second. In this book we use the word entity for those things that most people would call an 'object' because the term 'object orientation' has acquired a very special meaning in database technology and programming (see Chapter 3). In this jargon, 'object-orientation' is used to refer to a way of structuring data in the computer or in a computer program and does not necessarily mean that a physical entity is being referred to.

Continuous fields. In the continuous field approach, the simplest conceptual model represents geographical space in terms of continuous Cartesian coordinates in two or three dimensions (or four if time is included). The attribute is usually assumed to vary smoothly and continuously over that space. The attribute (e.g. air pressure, temperature, elevation above sea level, clay content of the soil) and its spatial variation is considered first; only when there are remarkable clusters of like attribute values in geographical space or time, as with hurricanes or mountain peaks, or 'significant events' will these zones be recognized as 'things' (e.g. Hurricane Caesar, the Matterhorn, the Gulf Stream, or the clay layer rich in the element Indium that is thought to date the asteroid impact that caused the demise of the dinosaurs).

Objects in a vector GIS may be counted, moved about, stacked, rotated, colored, labeled, cut, split, sliced, stuck together, viewed from different angles, shaded, inflated, shrunk, stored and retrieved, and in general, handled like a variety of everyday solid objects that bear no particular relationship to geography. (Couclelis 1992)

Opting for an entity model or a continuous field approach can be difficult when the entities can also be seen as sets of extreme attribute values clustered in geographical space. Should one recognize Switzerland, for example, as a land of individual mountain entities (Eiger, Matterhorn, etc.) or as a land in which the attribute 'elevation' demonstrates extreme variation? In practice, a pragmatic solution based on the aims of the user of the database must be made. The choice of conceptual model determines how information can later be derived. Opting for an entity approach to mountain peaks will provide an excellent basis for a system that records who climbed the mountain and when, but it will not provide information for computing the slopes of its sides. Choosing a continuous representation allows the calculation of slopes as the first derivative of the surface, but does not give names for those parts of the surface where the first derivative is zero and the curvature is in every direction downwards i.e. the peaks.

. . . the phenomenon of interest is blithely bisected by the image frame . . . for the mindless mechanical eye everything in the world is just another array of pixels. (Couclelis 1992)

As a gross oversimplification, the choice of an entity or a field approach also depends on the scientific

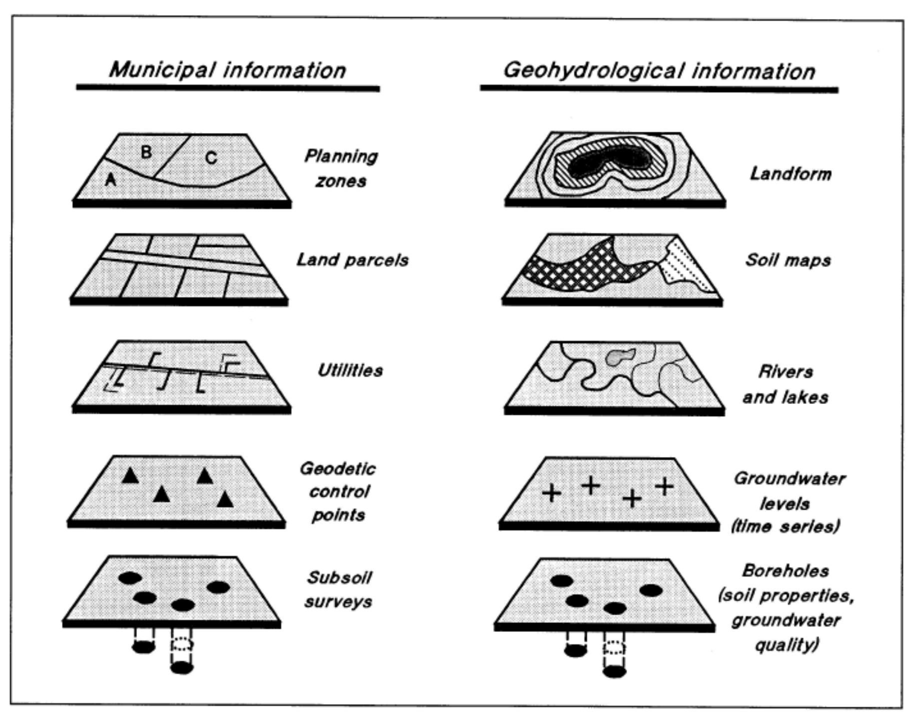


Figure 2.2. Examples of the different kinds of geographical data collected for different purposes by persons from different disciplines

or technical discipline of the observer. Disciplines that focus on the understanding of spatial processes in the natural environment may be more likely to use the continuous field approach while those who work entirely in an administrative context will view an area as a series of distinct units (Figure 2.2).

Geographical data models and geographical data primitives

Geographical data models are the formalized equivalents of the conceptual models used by people to perceive geographical phenomena (in this book we use the term 'data type' for the kind of number used to quantify the attributes—see below). They formalize how space is discretized into parts for analysis and communication and assume that phenomena can be uniquely identified, that attributes can be measured or specified and that geographical coordinates can be registered. As data may be collected in a variety of ways, information on the method or the level of resolution of observation or measurement may also be an important part of the data model.

Most anthropogenic phenomena (houses, land parcels, administrative units, roads, cables, pipelines, agricultural fields in Western agriculture) can be handled best using the entity approach. The simplest and most frequently used data model of reality is a basic spatial entity which is further specified by attributes and geographical location. This can be further

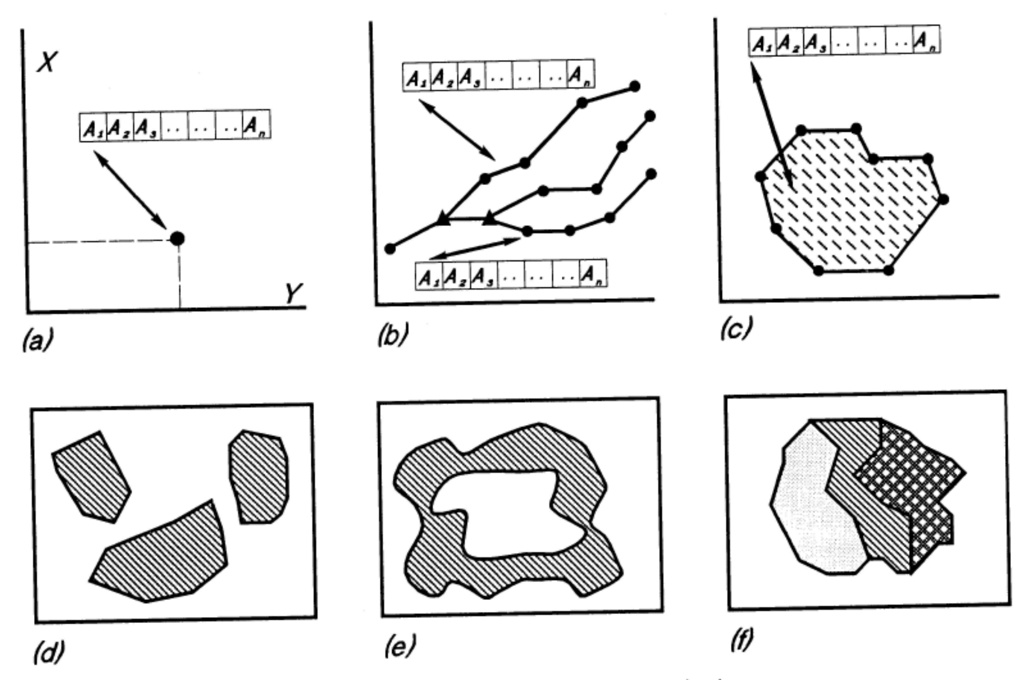


Figure 2.3. The fundamental geographical primitives of points, lines, and polygons

subdivided according to one of the three basic geographical data primitives, namely a 'point', a 'line', or an 'area' (which is most usually known as a 'polygon' in GIS) which are shown in Figure 2.3. These are the fundamental units of the vector data model and its various forms are summarized in Table 2.1 and illustrated in Figure 2.4*a*,*c*. Alternative means of representing entities using tessellations of regular-shaped polygons are to use sets of pixels (see below).

With continuous field data, although the variation of attributes such as elevation, air pressure, temperature, or clay content of the soil is assumed to be continuous in 2D or 3D space (and also in time), the variation is generally too complex to be captured by a simple mathematical function such as a polynomial equation. In some situations simple regression equations (trend surfaces) may be used to represent large-scale variations in terms of simple, differentiable numerical functions (see Chapter 5) but generally it is necessary to divide geographical space into discrete spatial units as given in Table 2.1 and shown in Figure 2.4*b*,*d*. The resulting tessellation is taken as a reasonable approximation of reality at the level of resolution under consideration and it is assumed that the operations such as differentiability which can be applied to continuous mathematical functions also apply to these discretized approximations.

Both the entity and tessellation models assume that the phenomena can be specified exactly in terms of both their attributes and spatial position. In practice there will be some situations where these data models are acceptable representations of reality, but there will be many others where uncertainties force us to choose pragmatically the one or the other approach (the effects of uncertainty and error in spatial analysis are dealt with in Chapters 9 and 10).

Vector data models of entities

The vector data model represents space as a series of discrete entity-defined point line or polygon units which are geographically referenced by Cartesian coordinates as shown in Figure 2.3.

Simple points, lines, and polygons Simple point, line, and polygon entities are essentially static representations of phenomena in terms of XY coordinates. They are supposed to be unchanging, and do not contain any information about temporal or spatial variability. A point entity implies that the geographical

Table 2.1. Discrete data models for spatial data

Vector representation of exact entities	Tessellations of continuous fields
Non-topological structures (loose points and lines 'spaghetti')	Regular triangular, square, or hexagonal grid (square pixels = raster)
Simple topology with linked lines—e.g. a drainage net or utility infrastructure	Irregular tesselation: Thiessen polygons
Complex topology with linked lines and nested structures—e.g. linked polygons	Triangular irregular nets (TIN)
Complex topology of object orientation with internal structures and relations	Finite elements
	Nested regular cells/quadtrees Irregular nesting

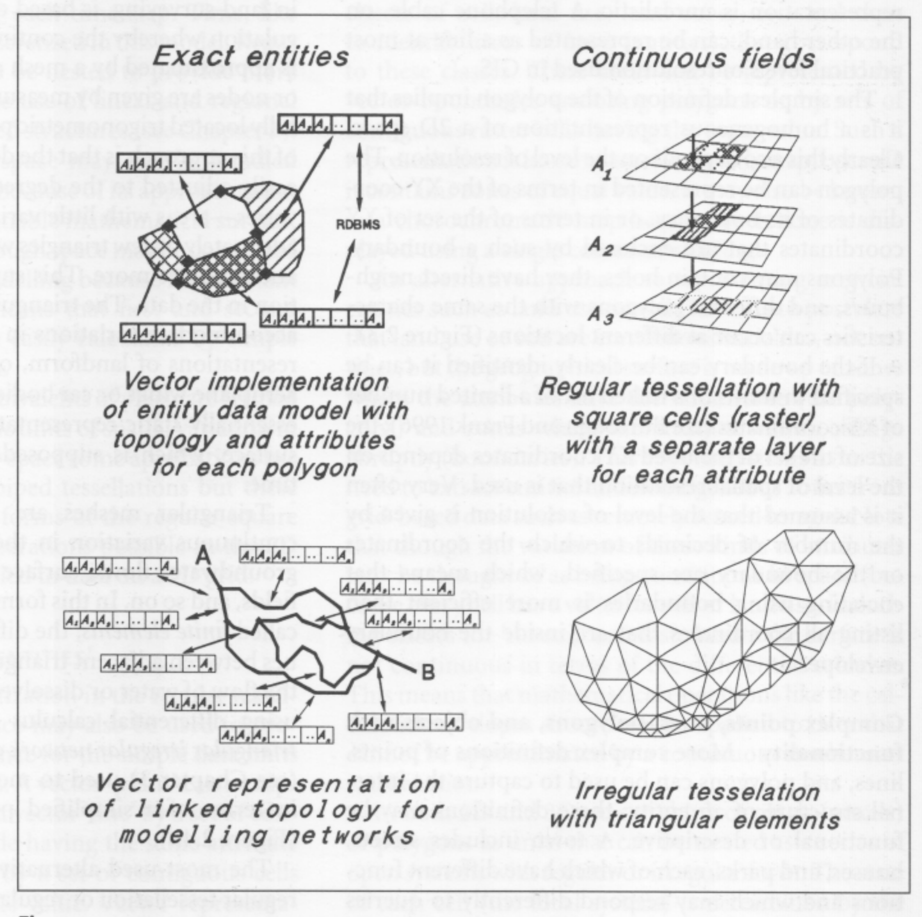


Figure 2.4. The encoding of exact objects (entities) and continuous fields in different data models. (*a*) top left: vector representation of crisp polygons; (*b*) top right—raster model of continuous fields; (*c*) bottom left—vector representation of linked lines; (*d*) bottom right—Delaunay triangulation of a continuous field

extents of the object are limited to a location that can be specified by one set of XY coordinates at the level of resolution of the abstraction. A town could be represented by a point entity at a continental level of resolution but as a polygon entity at a regional level. Increasing the level of resolution reveals internal structure in the phenomenon (in the case of a town, subdistricts, suburbs, streets, houses, lamp-posts, traffic signs) which may be important for some people and not for others.

A line entity implies that the geographical extents of the object may be adequately represented by sets of XY coordinate pairs that define a connected path through space, but one that has no true width unless specified in terms of an attached attribute. A road at national level is adequately represented by a line; at street level it becomes an area of paving and the line representation is unrealistic. A telephone cable, on the other hand, can be represented as a line at most practical levels of resolution used in GIS.

The simplest definition of the polygon implies that it is a homogeneous representation of a 2D space. Clearly this also depends on the level of resolution. The polygon can be represented in terms of the XY coordinates of its boundary, or in terms of the set of XY coordinates that are enclosed by such a boundary. Polygons may contain holes, they have direct neighbours, and different polygons with the same characteristics can occur at different locations (Figure 2.3).

If the boundary can be clearly identified it can be specified in terms of a linked list of a limited number of XY coordinates (see Burrough and Frank 1996); the size of the set of included XY coordinates depends on the level of spatial resolution that is used. Very often it is assumed that the level of resolution is given by the number of decimals to which the coordinates or the boundary are specified, which means that encoding using boundaries is more efficient than listing all coordinates that are inside the boundary envelope.

Complex points, lines, polygons, and objects with functionality More complex definitions of points, lines, and polygons can be used to capture the internal structure of an entity; these definitions may be functional or descriptive. A town includes streets, houses, and parks, each of which have different functions and which may respond differently to queries or operations at the town level. Topological links (Figure 2.4*a*,*c*) can be used to indicate how lines are linked into polygons or linked networks, respectively. In recent years more highly structured ways of encapsulating entity data have been possible through the object-oriented approach. This is the technical name for database and programming tools that provide nested hierarchies and functional relations between related groups of entities that together form a single unit at a higher aggregation level. Object orientation is described more fully in the context of data structures in Chapter 3.

Tessellations of continuous fields

Continuous surfaces can be discretized into sets of single basic units, such as square, triangular, or hexagonal cells, or into irregular triangles or polygons (the Thiessen/Dirichlet/Voronoi procedure—see Chapter 5) which are tessellated to form geographical representations. The use of irregular triangles, long used in land surveying, is based on the principle of triangulation whereby the continuous surface of the land is approximated by a mesh of triangles whose apices or nodes are given by measured 'spot heights' at carefully located trigonometric points. A major advantage of this approach is that the density of the mesh can be easily adjusted to the degree with which the surface varies—areas with little variation can be represented adequately by few triangles while areas with large variation require more. This supports a variable resolution in the data. The triangular surface can also easily accommodate variations in form as seen in 3D representations of landform, or other surfaces such as aeroplane wings or car bodies. These data models are essentially static representations of the hypsometric surface, which is supposed to be unchanging over time.

Triangular meshes are also used to represent continuous variation in the dynamic modelling of groundwater flows, surface water movement, wind fields, and so on. In this form they provide a structure called *finite elements*; the differences in attribute values between adjacent triangular cells that result from the flow of water or dissolved materials are modelled using differential calculus (Gee *et al.* 1990). The *triangular irregular network* or *TIN* is a data structure (see Chapter 3) used to model continuous surfaces in terms of a simplified polygonal vector schema (Figure 2.4*d*).

The most-used alternative to triangulation is the regular tessellation or regular grid. The 2D geometric surface is divided into square cells (known as *pixels*) whose size is determined by the resolution that is required to represent the variation of an attribute for a given purpose. The grid cell representation of space is

known as the *Raster* data model (Figure 2.4*b*). When the grid cells are used to represent the variation of a continuously varying attribute each cell will have a different value of the attribute; the variations between cells will be assumed to be mathematically continuous so that differential calculus may be used to compute local averages, rates of change, and so on. Each grid cell may be thought of as a separate entity that differs from vector polygons only in terms of its regular form and implicit rather than explicit delineation (Tobler 1995).

Although the regular grid is most often used to represent static phenomena, it can easily be adapted to deal with dynamic change. Changes over time may be recorded in separate layers of grid cells, one for each time step, so that the change from the static to the dynamic data model requires only that the basic structure is repeated for each time step. Time, like space, is assumed to be discretized in this model. Regular tessellations may also be nested to provide more spatial detail through the use of linear and regional quadtrees and other nested structures (see Chapter 3).

Lateral changes over space may also be handled easily by the regular grid because of its approximation to a continuous, differentiable mathematical surface. The flow of materials through space may be computed using *finite difference* modelling because the constant geometry of the cells means that first and second order derivatives can be easily calculated by simple subtraction and addition.

Three-dimensional equivalents of pixels are termed *voxels*—these are the basic units of spatial variation in a 3D space. In 2D and 3D space some applications use rectangular or parallelepiped tessellations but these can be seen as aberrant forms of the regular square discretization. All the operations possible in the 2D regular grids can be applied to data on a 3D grid.

Pixels and voxels as 'entities'

The basic units of discretization in the regular tessellation of continuous space may also be used to provide a geometrical reference for the simple data units of points, lines, and areas. A vector 'point' can be represented by a single cell; a vector 'line' by a set of contiguous cells one cell wide having the same attribute value; a vector polygon by a set of contiguous cells having the same attribute value. Vector representation is often preferred because regular grid cells may lose spatial details, though this is becoming less of a problem with increased power of computers and memories.

This equivalence between the vector and raster models of space frequently causes confusion about the nature of the phenomena being represented. For example, continuous fields can also be represented by isolines or contours, which are sets of XY coordinates linking sites of equal attribute value. Contours are useful ways of representing attribute values on paper maps and computer screens but they are less efficient for handling continuous spatial variation in numerical models of spatial interactions. Contour envelopes may be treated as simple polygons or as closed lines in terms of the entity approach, but they are merely artefacts of representation, not outlines of real world 'objects'.

The overlap between the two approaches is greatest when we have to deal with phenomena such as soil mapping units, or land use or land cover units. The classic approach of conventional mapping is to define classes of soil, land use, land cover, etc. and then to identify areas of land (entities) that correspond to these classes. These areas can be represented by vector boundaries enclosing polygons or by sets of contiguous raster cells having the same value. Such a representation is known as a choropleth map, because it contains zones of equal value. It may also be known as a chorochromatic map because each zone is displayed using a single colour or shading.

An alternative approach to representing artificial entities such as land use or soil classes is to postulate that land use or soil are continuous variables, not entities, but the geographical surface is made up of zones where the attributes have the same value (the polygons) and zones where the attribute values change abruptly (boundaries). This approach stems from the need to extract entities and homogeneous zones from grid-based data such as remotely sensed images. Note that though the variation of these kinds of attributes may be thought of as being continuous in space (because every cell has a value for soil or land use, including classes for 'no soil' or 'waste land') the surface is not continuous in terms of the differential calculus. This means that mathematical operations like the calculation of slopes should not be applied to data that cannot be approximated by a continuous mathematical function

In both 2D and 3D we can think of pixels (voxels) or polygons as units that can be treated as a series of open systems with regular or irregular form. The state of each cell (the local system) is determined by the value of its attributes; attribute values can be changed by operations that refer only to the cell in question or that use information from other cells that in one way or another are part of the cell's surroundings. The

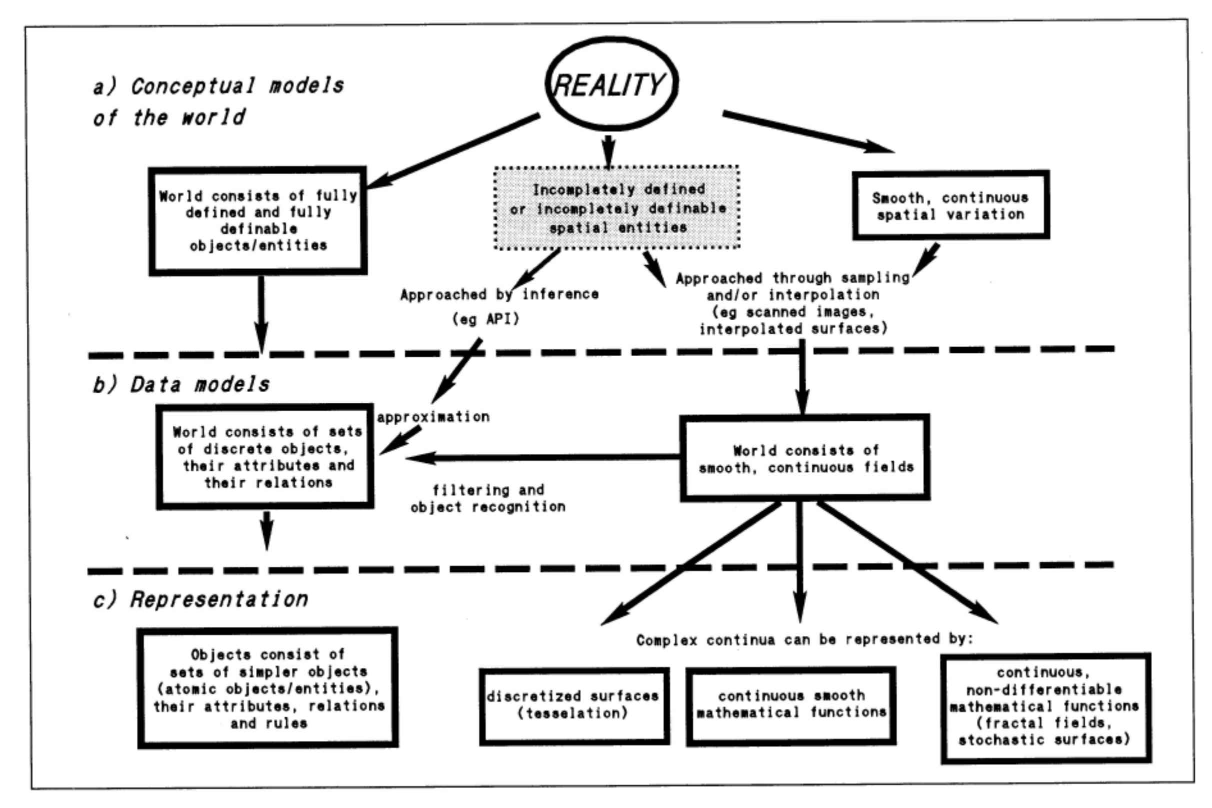


Figure 2.5. Steps in the process from observation of real word phenomena to the creation of standardized data models

difference between this approach for cells and the approach for polygons is not great (Tobler 1995); only the variable geometry of the vector representation of entities means that computing neighbourhood interactions is much more complex than with regular cell structures.

Figure 2.5 summarizes the data modelling stages followed so far.

The display of geographical primitives using vector and raster approaches

People are used to seeing spatial information represented both by lines or dotted shading (for example see the paintings of Seurat or use a lens to see the dots that make up a photograph in the newspaper). If we look at the conventional way of communicating geographical data, paper maps, these use both vector and continuous field data models in characterizing an area. The real world is portrayed either in terms of a series of entities represented by coloured or stylized point, line, or area symbols, or as a continuous variation in the values of an attribute over space, such as the portrayal of the elevation by the *hypsometric curve* or as contours.

In the computer these essentially linear or dotted approaches are formalized into the vector and raster methods of representation. Figure 2.6 presents the main ways in which the simple geographical data

VECTOR	Points	Lines	Areas
Feature data			a b d c a
Areal units		324 326	P24 P25
Networks			
Sampling records			f3 f4
Surface data			
Label/ text	Utrecht Arnhem	A12	Lake
Symbols			
Relations	attributes and pointers	attributes and pointers	

RASTER	Points	Lines	Areas
Feature data			
Areal units		-	
Networks	-	-	-
Sampling records		-	
Surface data		-	
Label/ text	-	-	-
Symbols			
Relations	attributes and relations	attributes and relations	

Figure 2.6. The different ways of graphically displaying data encapsulated by (*a*) left—vector entity models, and (*b*) right—raster models

models can be visualized in the vector or raster domain. The figure also summarizes and makes explicit the influence of cartographic semiology in the vector representation.

Data types

In everyday speech we distinguish qualitative or nominal attributes from quantitative data or numbers, and recognize that different kinds of operations suit different kinds of data. The same is true when describing geographical phenomena using a formalized data model and the information may be written down (and stored in the computer) using various *data types* (Table 2.2).

The *attributes* of entities may be expressed by Boolean, nominal, ordinal, integer, or real data types. Real data types include decimals; integer and real are collectively known as *scalar* data. Geographical *coordinates* are sometimes expressed as integers but mostly as real data types, and topological linkages use integers. *Differentiable continuous surfaces* require real data types though integers are sometimes used as an

Table 2.2. Data types

Data type	Allowed values	Allowed operations
Boolean	0 or 1	Logical and indicator operations: Truth versus Falsehood
Nominal	Any names	Logical operations, classification and identification
Ordinal	Numbers from 0 to ∞	Logical and ranking operations, comparisons of magnitude
Integer	Whole numbers from −∞ to +∞	Logical operations, integer arithmetic
Real	Real numbers (with decimals) from −∞ to +∞	All logical and numerical operations
Topological	Whole numbers	Indicate links between entities

approximation (one cannot take the derivative of a nominal attribute). *Non-differentiable continuous surfaces* and their discretized forms (grid cells or pixels) can take the same range of data types as entities.

Logical operations can be carried out with all data types, but *arithmetical operations* are limited to real and integer data types. Consequently the kind of data analysis is governed by the data types used in the data model. The limits of accuracy of arithmetical operations is limited by the length of the computer word used to record the numbers (see Chapter 8).

Axioms and procedures for handling data in information systems

Having explored the problems of defining data models as representations of phenomena in the real world and the ways they can be constructed from geographical primitives as simple or complex entities or discretized continuous surfaces we can now specify the logical ground rules and axioms (*sensu* 'generally accepted propositions or principles sanctioned by experience': *Collins English Dictionary*, 3rd edn., 1994) that govern the way these data models may be treated. Though some of the following may seem self-evident, the following statements (adapted from Robinove 1986) provide a formal basis for spatial data handling and look forward to the material presented in Chapters 7 and 8.

1. It is necessary to identify some kind of discretization such as *entities* (individuals) that carry the data. In GIS the primitive entities are *points*, *lines*, *polygons*, and *pixels* (grid elements). Complex entities having a defined internal structure can be built from sets of points, lines, and polygons.

2. All fundamental entities are defined in terms of their *geographical location* (spatial coordinates or geometry), their *attributes* (properties) and *relationships* (topology). These relationships may be purely geometrical (with respect to spatial relations or neighbours), or hierarchical (with respect to attributes) or both.

3. Individuals (*entities*) are distinguishable from one another by their attributes, by their location, or by their internal or external relationships. In the simple, static view, individuals or 'objects' are usually assumed to be internally homogeneous unless they are a representation of a mathematical surface or unless they are complex objects built out of the primitives. In most

cases GIS usually only distinguish objects that are internally homogeneous and that are delineated by crisp boundaries. A more complex GIS allows intelligence about inexactness in objects in which either the attributes, the relationships, or the location and delineation are subject to uncertainty or error.

4. Both entities and attributes can be classified into useful categories.

5. The propositional calculus (Boolean algebra) can be used to perform logical operations on an entity, its attributes, its relations, and the groups to which it belongs.

6. In GIS the propositional calculus is extended to take account of:

distance
direction
connectivity (topology)
adjacency
proximity
superposition
group membership
ownership of other entities

Intelligent GIS extend the propositional calculus to take account of non-exact data (see Chapter 11).

7. New entities (or sets of entities) can be created by geometrical union or intersection of existing entities (line intersection, polygon overlay)—see Chapter 6.

8. New complex entities or objects can be created from the basic point, line, area or pixel entities.

9. New attributes can be derived from existing attributes by means of logical and/or mathematical procedures or models.

$$\text{New attribute } U = f(A, B, C, \ldots)$$

The mathematical operations include all kinds of arithmetic (addition, subtraction, multiplication, division, trigonometry, differentiation and integration, etc.) depending on data type—see Table 2.2.

New attributes can also be derived from existing topological relations and from geometric properties (e.g. linked to, or size, shape, area, perimeter) or by interpolation.

10. Entities having certain defined sets of attributes may be kept in separate subdata sets called data planes or overlays.

11. Data at the same XYZt coordinate can be linked to all data planes (the principle of the common basis of location).

12. Data linked to any single XYZt coordinate may refer only to an individual at that coordinate, or to the whole of an individual in or on which that point is located.

13. New attribute values at any XYZt location can be derived from a function of the surroundings (e.g. computation of slope, aspect, connectivity).

Data modelling and spatial analysis

As should be clear there are direct links between these fundamental axioms, the data model, and the data type used to represent a geographical phenomenon, and the kinds of analysis that can be carried out with it. The following different situations help to illustrate this:

1. If the location and form of the entity is unchanging and needs to be known accurately, but the attributes can change to reflect differences in its state caused by inputs of new data or output from a numerical model, then the vector representation of the entity model is appropriate. This is the most common situation in conventional GIS.

2. If the attributes are fixed, but the entity may change form or shape but not position, as in the drying up of a lake, then a vector model requires a redefinition of the boundary every time the area of the lake changes. A raster model of a continuous field, however, would treat the variation of the water surface as a response surface to a driving process so that the extents of the lake could be followed continuously.

3. If the attributes can vary and the entity can change position but not form, or its separate parts are linked together, the behaviour can be well described by an object-oriented model which can pass information from one level of the model to another.

4. If no clear entities can be discerned, then it is often preferable to treat the phenomenon as a discretized, continuous field.

Examples of the use of data models

Collectors and users of geographical data need to make decisions on the choice of data model every day. The following examples illustrate the importance of an understanding of these ideas.

CADASTRE

The main aim of the cadastre or land registry is to provide a record of the division and ownership of land. The important issues are the location, area, and extent of the land in question and its attributes (such as the name and address of the owner), the address of the parcel in question, and information about transactions and legal matters. In this case the exact entity (vector) model works well, using nominal, integer, and real data types to record the attributes and real data types for the coordinates. In many countries land registry is highly organized, parcel boundaries are surveyed with high accuracy to reduce the chance of disputes so the assumption of a data model in which the parcel is bound by infinitely thin lines is a good approximation to what the land registry is trying to achieve. The coordinates of these boundaries are located accurately with respect to a national or local reference and the attributes of the entity in the database are simply the properties associated with the parcel. An essential aspect of the polygon representation is that boundaries may be shared by adjacent parcels. This saves double representation in the database (see Chapter 3) and links the boundaries into a topologically sound polygon net which can handle both adjacency and inclusions.

UTILITY NETWORKS

Utility network is the generic term for the collections of pipes and wires that link the houses of consumers to the supplies of water, gas, electricity, telephone, cable television to national or regional suppliers and also to the waste water disposal systems of drains and sewers. In many countries these networks are hidden below ground though in some, electricity, telephone, and television cables may be suspended from poles along the streets. Three aspects of these networks are important when recording these phenomena, namely (*a*) the attributes of a given net (what it carries, what kind of wire or pipe, additional information on materials used, age, name of contractor who installed it, and so on); (*b*) the location of the net (so that persons digging in the street will not damage it and so it can be found quickly when needing repair), and (*c*) information on how different parts of the net are connected together. Clearly all these requirements can be incorporated in a data model of topologically connected lines (entities) that are described by attributes (Figure 2.2). Data types may include all forms.

LAND COVER DATABASES

National and international governments are interested in the division of the landscape according to classes of land cover—urban areas, arable crops, grassland, forest, waterbodies, coasts, mountains, etc. Creating a data model for such an application requires several steps. First, it is necessary to define exactly what is meant by the classes. Second, one needs to decide how to recognize them, and third, one must choose a survey methodology (such as a point sample survey or a remote sensing scanner in a satellite) to collect surrogate data which are then interpreted to produce the result desired.

The simplest data model assumes that the classes are crisp and mutually exclusive and that there is a direct relation between the class and its location on the ground. If this is acceptable then one can use the simple polygon primitives as a model for each occurrence of each class. The result is the well-known choropleth, or more correctly, chorochromatic map. The issues involved in building the database are then defining the classes, identifying and mapping the boundaries, and attaching the attributes to each class in a manner that is equivalent to the polygon net model for cadastral mapping. The data types used will range from nominal (for recording names of classes) to scalar (for computing and recording areas).

Consider now what might happen if there are disagreements about how to classify land cover. Different people or organizations might have different reasons for allocating land to different classes; even if the same number of classes are used and the central concepts of the classes are similar. Table 2.3 presents some examples of how different the results can be. Clearly, any analyses based on data from the different sources would give considerably different results, with serious implications for interpretation.

Now consider the effects of the method of survey on the results. If we collect land cover data by sam-

Table 2.3. Examples of variation in estimation of land cover in Europe—km²*1000

Land use classification	FAO-Agrostat	Pan-European Questionnaire by Eurostat	10 minutes Pan-European Land Use Database	Land Use Statistical Database	Land Use Vector Database
Permanent crops					
Germany	4.42	—	1.80	2.30	5.36
France	12.11	—	12.07	12.18	31.45
Netherlands	0.29	—	0.22	0.34	1.07
UK	0.51	—	0.52	0.59	6.54
Forest					
Germany	103.84	103.84	98.56	—	100.46
France	147.84	148.10	140.675	145.81	79.63
Netherlands	3.00	3.30	1.48	3.00	0.78
UK	23.64	24.00	18.96	14.29	10.03

Source: RIVM 1994.

pling, i.e. by visiting a set of data points on the ground and recording what is there we will have to interpolate from these 'ground truth' data to all sites where no observations have been made. The allocation of an unsampled site to a given class is then a function of the quality and density of the data and the power of the technique used for interpolation. Note that if we decide to interpolate to crisply delineated areas of land we still use the choropleth model based on the geographical primitive area/polygon. If we interpolate to a discretized surface such as a regular grid then the land cover map consists of sets of pixels with attributes indicating the land cover class to which they belong.

If we use remotely sensed data to identify land cover we automatically work with a gridded discretization of continuous space because that is how the satellite scanner works. The resolution of our spatial information is limited by the spatial resolution of the scanner, which for digital orthophotos may be very fine indeed (Plate 1). Unlike the case with sampling, we do have complete cover of the area (excluding problems with cloud cover on the image and so on) so the information present in each pixel is of equal quality, which is not the case with interpolation. The major problem with identifying land cover with remotely sensed data is to convert the measurements of reflected radiation for each pixel into a prediction that a given land cover class dominates that cell. Obviously the success of the quality of the data depends on the quality of the classification process.

This example shows that two different, but complementary data models can be used for land cover mapping. The representation of these models as vectors or rasters depends partly on how the data have been collected and partly on the way they will be used.

Soil maps

Most published soil maps use the entity data model based on the vector polygon as the geographical primitive. Polygons are defined in terms of their soil class, which by implication is homogeneous over the unit. Boundaries are represented as infinitely thin lines implying abrupt changes of soil over very short distances. Note that inclusion of polygons in polygons is an important aspect of soil and geological maps.

This data model is practical, because it means that simply by locating a site on a map and determining the mapping unit one can retrieve information on the soil properties by consulting the survey report. However, the paradigm is scientifically inadequate because it ignores spatial variation in both soil forming processes and in the resulting soils (see Plates 4.4, 4.6). The model is conceptually identical to that used in other polygon-based spatial information, such as parcel-based land use or land ownership data, and it has provided the role model for the development of soil information systems and GIS in the late 1970s and 1980s (Burrough 1991*b*). Data types used to record attribute data will include nominal, integer, and real.

A critical aspect of delineating soil classes in geographical space concerns the interpretation of boundaries on the ground. Important soil differences may be indicated by abrupt, clearly observable physiographic features such as changes in lithology, drainage, or breaks of slope, which we can call 'primary boundaries'. However, the drawn soil boundaries may also merely reflect interpreted differences in soil classification in the data space. These we term 'secondary boundaries'.

As far as the user is concerned, printed (and digitized) soil maps do not distinguish between 'real' primary boundaries that have a physiographic basis, and 'interpreted' secondary boundaries. Consequently, as both types of boundary are represented as supposedly infinitely thin lines, the mapping procedures lead inevitably to a 'double crisp' conceptual model of soil variation, both with respect to the classification in attribute space and the geographical delineation of mapping units. According to this Boolean model any site can belong to a only single soil unit: both in attribute space and geographical space the membership of a site in soil class i is either 0 (not a member or outside the area) or 1 (is a member or is inside).

A major problem with soil, vegetation, and other similar natural phenomena, is that they vary spatially at all scales from millimetres to whole continents. Although soil scientists have long recognized this, soil cartographers still use the choropleth model for mapping soil at different levels of resolution. This generates a serious logical fallacy because at one level it assumes within-unit homogeneity, while at another spatially coherent differences in soil have been recognized. In addition, whereas with land cover we might expect the boundaries of land cover classes to be reasonably sharp and discrete because their location is often dictated by differences in human use of landscape, real soil boundaries can be sharp, gradual, or diffuse.

An alternative to the discrete polygon data model for soil is to assume that soil properties vary gradually over the landscape. The soil is sampled at a series of locations and attributes are determined for these samples. The simplest data model is then the geographical point (to represent the locations) with the values of the associated attributes. From this simple data model new data models of continuous spatial variation may be created by interpolation.

The data models for describing soil as a continuous variable are, in principle, very similar to those used for the hypsometric surface of land elevation. Sets of discrete contour lines can be used to link zones of equal attribute values, or attribute values may be interpolated to cells or locations on a regular grid, which leads to the raster model of space. As with soil and many other attributes of the physical, chemical, and biological landscape, these cannot be seen directly but must be collected at sets of sample points according to some approved sampling scheme. Both the area (or volume) of the samples (known technically as the *support*) and their density in space relative to the spatial variation of the attribute concerned are important for the quality of the resulting interpolations. The details of interpolation and the differences in results obtained using different methods are explained in Chapters 5 and 6.

Hydrology

Hydrological applications require the modelling of the transport of water and materials over space and time, which can require changes to be signalled in attributes, and in location and form of critical patterns (e.g. water bodies). Not only may water levels in rivers, reservoirs, and lakes change but the geometry and location of water bodies can vary as well. A change in water level in a lake causes the location of the boundary between water and land to change. A flood may result in the opening up of new channels and the abandonment of old ones, thereby changing both topology and location.

The simple entity vector data model of points, lines, and areas is not very well suited to dealing with hydrological phenomena because changes in geometry mean changing the coordinate and topological data in polygon networks, which can involve considerable computation. Better is to use a data model based on ideas of 'object orientation' in which primitive entities are linked together in functional groups (McDonnell 1996). The internal structure of the data model permits action on one component of the group to be passed automatically to other parts; consequently the data model contains not only geographical location, geometry, topology, and attributes but also information on how all these react to change.

Transport of material can also be easily captured by data models of continuous variation. The use of the variable resolution triangular or square finite element net is common in hydrological models but not in commercial GIS (e.g. McDonald and Harbaugh 1988). Transport of material over a surface can also be dealt with using the raster data model of continuous variation to which the surface topology has been added or derived by computation (see Chapter 8).

Summary: entities or fields?

Figure 2.5 summarizes the steps that need to be taken when going from a perception of reality to a set of data models that can be used in a computerized information system. In some applications the decision to opt for an entity-based approach or a field-based approach may be clear-cut. In others it may be a matter of opinion depending on the aims of the user.

Interconversion between an entity-based vector or raster representation and a continuous representation is technically possible if the original phenomena have been clearly identified. Raster–Vector conversion is covered in Chapter 4. No amount of technology, however, can make up for differences in interpretation that are made before the phenomena are recorded. If scientist X perceives the landscape as being made up of sets of crisp entities represented by polygons, his view of the world is functionally different from scientist Y, who prefers to think in terms of continuous variation. Both approaches may be distortions of a complex reality which cannot be described completely by either model.

Figure 2.2 illustrates how the preferred choice of entities or continuous fields may vary between applications and within disciplines. Generally speaking, those disciplines concerned with the inventory and recording of static aspects of the landscape opt for the entity approach; disciplines dealing with the studies of pattern and process requiring dynamic data models opt for continuous, differentiable fields.

In most GIS all locational data and attribute values are deemed to be exact. Everything is supposed to be known, there is no room for uncertainty. Very often this comes not because we are unable to cope with statistical uncertainty, but rather because it costs too much to collect or process the data needed to give us the information about the error bands that should be associated with each attribute for each data unit. But, in principle there is no reason why information on quality cannot be added to the data.

In essence, the intellectual level of the simple crisp entity models of spatial phenomena is little different from that of children's plastic building blocks. These toys obey the basic axioms of information systems, including that which says that it is possible to create a wide variety (possibly an infinite variety?) of derived objects by combining various blocks in different ways. Logically this is no different from combining sets of points, lines, and areas from a GIS to make a new map. Given enough bricks one can build houses, recreate landscapes, or even construct life-sized models of animals like giraffes and elephants.

And this is the point. The giraffe built out of plastic blocks can be the size, the colour, and the shape of a real giraffe, but the model does not, and cannot have the functions of a giraffe. It cannot walk, eat, sleep, procreate, or breathe because the basic units of which it is built (the blocks or database elements) are not capable of supporting these functions. While this is a trivial example, the same point can be made for many database units that are used to supply geographical data to drive analytical or process oriented models. No amount of data processing can provide true functionality unless the basic units of the data models have been properly selected.

Nine factors to consider when embarking on spatial analysis with a GIS

The following nine, not necessarily independent, questions concerning spatial data are of fundamental importance when choosing data models, and database approaches for any given application:

1. Is the real world situation/phenomena under study simple or complex?
2. Are the kinds of entities used to describe the situation/phenomena detailed or generalized?
3. Is the data type used to record attributes Boolean, nominal, ordinal, integer, real, or topological?
4. Do the entities in the database represent objects that can be described exactly, or are these objects complex and possibly somewhat vague? Are their properties exact, deterministic, or stochastic?
5. Do the database entities represent discrete physical things or continuous fields?
6. Are the attributes of database entities obtained by complete enumeration or by sampling?
7. Will the database be used for descriptive, administrative, or analytical purposes?
8. Will the users require logical, empirical, or process-based models to derive new information from the database and hence make inferences about the real world?
9. Is the process under consideration static or dynamic?

Questions

1. Develop simple data models for use in the following applications:

   A road transport information system
   The location of fast food restaurants
   The incidence of landslides in mountainous terrain
   The dispersion of pollutants in groundwater
   An emergency unit (police, fire, ambulance)
   A tourist information system
   The monitoring of vegetation change in upland areas
   The monitoring of movement of airborne pollutants, such as the ^{137}Cs deposited by rain from the Chernobyl accident in 1986.

Consider the sources of data, the kinds of phenomena being represented, the data models, the data types, and the main requirements of the users.

Suggestions for further reading

COUCLELIS, H. (1992). People manipulate objects (but cultivate fields): beyond the raster–vector debate in GIS. In A. U. Frank, I. Campari, and U. Formentini (eds.), *Theories and Methods of Spatio-Temporal Reasoning in Geographic Space*. Springer Verlag, Berlin, pp. 65–77.

GAHEGAN, M. (1996). Specifying the transformations within and between geographic data models. *Transactions in GIS*, 1: 137–52.

GATRELL, A. C. (1991). Concepts of Space and Geographical Data. In D. J. Maguire, M. F. Goodchild, and D. W. Rhind (eds.), *Geographical Information Systems: Principles and Applications*, vol. ii. Longman Scientific and Technical, Harlow, pp. 119–34.

HARVEY, D. (1969). *Explanation in Geography*. Edward Arnold, London.

THREE

Geographical Data in the Computer

To take advantage of the possibilities a computerized spatial modelling tool may provide, it is essential to understand how the data models used to represent geographical phenomena are coded in these devices. This chapter describes the ways in which spatial data may be efficiently coded into a computer to support the operations of a GIS. Computers code data and carry out instructions using a series of switches which exist in one of two states 'on' or 'off'. These states are coded by the numbers 1 and 0 respectively and the base 2 binary system is therefore the fundamental coding basis for all computing. Therefore geographical data need to be converted into discrete records in the computer using these switches to represent the location, presence or absence, type etc. of the phenomenon. The main data structures used to date are termed vector and raster and they divide space into either a series of points, lines, and areas, or into a regular tessellation. Recently the object-oriented data structure has been used which employs the same basic units as the vector and raster systems but structures the entities into linked or hierarchical forms.

When entered into the computer the data need to be organized to allow efficient accessing, retrieval, and manipulation. Systems for organizing data in the computer range from simple lists or indexed files, through to highly structured databases based on hierarchical, network, relational, and object-oriented schemata. These structures determine how the data are organized into data records and so control the way they are held in the computer. The various combinations impose limitations on the data representation and handling which affect the analysis and modelling of the information. Different ways of coding and storing the various data structures in a computer are discussed. In recent years research has highlighted the possibilities of using object-oriented database structures in GIS which support a more realistic structuring for entity data.

A computer version of a map may be no more than a digital image of a paper document stored on an electronic storage device such as a CD-ROM, or it may be an interactive display of information held in a complex GIS. The latter form of map may have been made using procedures that are different from conventional topographical or thematic maps and may use other kinds of symbolism to convey its message so it is essential to understand how the end-product relates to the original observations of the phenomena in question.

The computer provides a much greater range of possibilities for recording, storing, retrieving, analysing, and displaying spatial data than a conventional map. Whereas the conventional paper map was a general purpose database created by a survey at a given point in time, the computer can select any view of the data for any given purpose. The computer allows us to add data, to retrieve data, and to compute new information from existing data, so a computerized spatial database need not remain static, but may be used to model changes in spatial phenomena over time. These new possibilities raise the following questions about the nature of space, on spatial data, and how spatial phenomena should be described in terms of conceptual models.

If our aim is to automate the map-making process, to provide an electronic analogue of the conventional map (in which the graphic representation of spatial phenomena is coded by standard symbols and shading) then the conceptual models required for computerizing the data differ little from those used in conventional mapping. The main task is to devise ways of coding the formalized geographical data models so that the computer can handle them. The links between conceptualization–data storage–data retrieval–visualization are similar to those used in conventional map-making, with the computer doing the hard work of retrieving and drawing the maps.

If our aim, however, is to extend our abilities to handle spatial data in such a way that our view of the world is not limited by the constraints of the printed map, if we wish to extract subsets of spatial data for specific purposes, or to link information on spatial phenomena to the physical processes governing their appearance and distribution, then the computer provides many opportunities beyond the automation of existing tasks. A computer provides the means to interact with data in ways that are impossible with printed information. Data can be changed, retrieved, recomputed, and displayed not just as a reflection of new information that has been collected from the real world, but also in response to numerical models of spatial and temporal processes. The GIS is then more than a simple automator of existing tasks; it provides both an archive of spatial data in digital form and a tool for exploring the interactions between process and pattern in spatial and temporal phenomena.

When geographical data are entered into a computer it would be very convenient if the GIS recognized the same conceptual models the user is accustomed to (discussed in Chapter 2). This would allow the interactions between the user and the computer to be both possible and intuitive. However, human perception of space is frequently not the most efficient way to structure a computer database and does not account for the physical requirements of storing and repeatedly using digital information. Computers for handling geographical data therefore need to be programmed to represent the phenomenological structures in an appropriate manner, as well as allowing the information to be written and stored on magnetic devices in an addressable way. Unlike many other kinds of data handled routinely by modern information systems, geographical data are complicated because they refer to the position, attributes, and the internal and external topological connections of the phenomena recorded. The topological and spatial aspects of geographical data distinguish them from the administrative records handled by data processing systems used for banking, libraries, airline bookings, or medical records. In addition GIS must display the data using graphical components and form, such as colour, symbols, shading, and line type and width, that are interpretable by people. Therefore as a result of the nature, volume, and complexity of geographical data a number of logical computer schemata have been developed to pack data onto devices for efficient storage, updating, and retrieval. These schemata are a continuance of the model/structure development processes discussed in Chapter 2 and are stages (*d*) and (*e*) of Box 2.1 (reproduced below for your convenience).

Once data have been stored in the database, they can be retrieved and transformed, again according to formalized logical and mathematical rules. New data may be created from old; data can be used over and over again without deteriorating in quality, though their intrinsic quality may affect the quality of the results—Chapters 6, 7, 8, 9, and 10 cover these aspects of geoinformation processing. First, it is important to consider how computers store data.

Spatial data models and data structures

The creation of analogue and digital spatial data sets involves seven levels of model development and abstraction (cf. Peuquet 1984*a*, Rhind and Green 1988, Worboys 1995):

(*a*) A view of reality (conceptual model)
(*b*) Human conceptualization leading to an analogue abstraction (analogue model)
(*c*) A formalization of the analogue abstraction without any conventions or restrictions on implementation (spatial data model)
(*d*) A representation of the data model that reflects how the data are recorded in the computer (database model)
(*e*) A file structure, which is the particular representation of the data structure in the computer memory (physical computational model)
(*f*) Accepted axioms and rules for handling the data (data manipulation model)
(*g*) Accepted rules and procedures for displaying and presenting spatial data to people (graphical model)

Data in the computer

People represent numbers and do arithmetic using the decimal system with the base 10. Each position in the written form of a number indicates how many units of the power of 10 correspond to that position. So the first column to the left of the decimal represents multiples of unity (10^0), the second multiples of 10 (10^1), the third multiples of 100 (10^2), and so on. Computers use a different counting system. They store data in the form of arrays of switches that have only two states; they are 'on' or they are 'off' and these states are coded by the numbers 1 and 0 respectively. Therefore computers code data and do computations using *binary arithmetic* which is base 2. Box 3.1 demonstrates the binary system and arithmetic. In the following discussion $_{10}$ is used to indicate a number in the decimal system and $_2$ in the binary system.

Binary numbers and computer representation of numbers and text

As with ordinary numbers, the data in the binary system are represented using a series of 'columns' and in a computer each column is used to represent a switch. The switches, which are really small magnetized areas on a computer disk or memory, are usually grouped in packets of eight, known as a *byte*. Several bytes can be linked together to make a *computer word*. Words are grouped together in a *data record* and records are grouped together in a *computer file*. Sets of computer files can be grouped together hierarchically in *directories* or *subdirectories*.

With a byte of eight bits we can represent 256 numbers from 0 to ($2^8 - 1$) i.e. from 0 to 255. For example the sequence of eight switches 11111111_2 means:

$$2^7*1 + 2^6*1 + 2^5*1 + 2^4*1 + 2^3*1 + 2^2*1 + 2^1*1 + 2^0*1 = 255_{10}$$

Remember that any positive number raised to a zero power equals 1.

If we combine 2 bytes in a 16-bit word it is possible to code numbers from 0 to 65 535. However, it is also useful to be able to code positive and negative numbers so only the first fifteen bits (counting from the right) are used for coding the number and the sixteenth bit (2^{15}) is used to determine the sign. This means that with sixteen bits we can code numbers from –32 767 to +32 767. A number lacking a fractional component is called an *integer*, so numbers in the range –32 767 to +32 767 are called *16-bit integers*.

Frequently it is necessary to code numbers that are larger or smaller than –32 767 to +32 767, or numbers that have fractional components, so computer

BOX 3.1.

Binary numbers and arithmetic

The binary system is in base 2 where numbers count using 1 and 0. As with ordinary numbers (the decimal system), the data are represented using a series of 'columns' where the first counts in units of 2^0 (i.e. 0 or 1), the second counts in units of 2^1 (0 or 2), the third counts in units of 2^2 (0 or 4) and so on

The binary sequence 0010 means $2^3 * 0 + 2^2 * 0 + 2^1 * 1 + 2^0 * 0 = 2_{10}$

The binary sequence 1100 means $2^3 * 1 + 2^2 * 1 + 2^1 * 0 + 2^0 * 0 = 12_{10}$

Binary arithmetic is similar to decimal arithmetic but there are some differences.

Addition and subtraction are similar:

$14_{10} - 3_{10} = 11_{10}$ is $1110_2 - 0011_2 = 1011_2$

Multiplication proceeds by adding numbers together

$23_{10} * 4_{10}$ is carried out as 23 + 23 + 23 + 23 = 92

the binary equivalent is $10111_2 + 10111_2 + 10111_2 + 10111_2 = 1011100_2$

Division proceeds by subtracting numbers (the reverse of the example above)

$1011100_2/10111_2 = 100_2$ i.e. 4_{10}

Multiplying by 2 is quickly achieved by shifting the columns one place to the left

2 * 0101 is 1010

Division by 2 is shifting the columns one place to the right

1110/2 = 0111

Of course, if the right most bit is a '1', division results in it being removed from the word so that simply multiplying by 2 (the left shift) again does not give the number one started with unless other checks are added.

words with more bits are needed. Real numbers with positive and negative values and decimals require 32-bit or even 64-bit words (so-called *double precision*) in which some of the bits are reserved for the decimal part (the part smaller than 1) and the rest are used for larger numbers. This method of coding numbers means that the accuracy with which any number may be coded is a function of the number of bits used. This means numbers are not coded exactly, which may lead to rounding errors when the user attempts to work with numbers that are larger than those allowed. Chapter 9 discusses the problems of rounding errors in GIS in more detail.

Data organization and coding using hexadecimals

So far we have considered the coding of decimal numbers into machine usable form but computers are also used to store text characters. One of the most commonly used systems for coding alphanumeric data is known as the American Standard Computer Information Index (ASCII) and is based on the *hexadecimal system*.

This system works with units of 4 bits which is equivalent to counting to the base 16 (2^4). Hexadecimal numbers may be represented by single characters

(as in the series 0–9) by replacing the numbers 10, 11, 12, 13, 14, 15 by the symbols A, B, C, D, E, F respectively. The column/place system is used just as with binary or decimal numbers. For example:

the decimal number 17 is hexadecimal 11 and binary 10001;
the decimal number 182 is hexadecimal B6 and binary 10110110.

The hexadecimal system also provides a convenient basis for coding alphanumeric data such as the letters of the alphabet and the numbers from 0 to 9 using a byte of 8 bits to code for a written symbol. So ASCII codes for text characters can also be written as hexadecimal numbers from zero to FF (decimal 255 or 2^8). For example:

the letter 'A' is coded as 41 hex (decimal 065, binary 01000001);
the number character '5' is coded as 35 hex (decimal 053, binary 00110101)
and the Greek letter 'μ' is coded as 82 hex (decimal 130, binary 10000010).

As both numbers and codes for text characters are represented by strings of bits that are set to 0 or 1, additional instructions need to be added to the computer file to indicate whether the information is coded directly or as a representation of a printed text.

The ASCII system is by no means the only way of coding text and many other systems are not only possible, but have been devised for coding text characters, diagrams, images, sound, and other information. The organization of the bits is known as its *format*; clearly unless the format is known, information cannot be extracted from the string of bits.

Data volumes

The volumes of bits needed to encode information are expressed in thousands (kilo-), millions (mega-), or billions (giga- or tera-) of bytes. A standard A4 page coded in ASCII format with 64 lines of text having 80 characters per line (including spaces) requires 40 960 bits (5 120 bytes or 5 k (1 k = 1024 bytes); a book of 200 pages requires 1.024 Mbyte for the characters alone, and additional space is required for the information on page numbers, text fonts, layout, etc. Maps and other forms of spatial data require large amounts of storage space.

From the previous section it follows that the relationship between data type and the number of bits needed to code the data means that it is possible to save computer space and processing time when dealing with one kind of data as compared to another.

Coding the basic data models for input to the computer

The operations of *data input* are the conversion of information from a form that people have perceived or can recognize, or as recorded on an automatic device, to a form that is suitable for numerical processing in the computer. In the previous chapter we saw how it is necessary to define formally conceptual data models based on geographical primitives to represent the data. Chapter 4 deals with the practical aspects of building a database and creating useful output; this chapter is concerned with the ways in which spatial data may be efficiently coded to carry out the tasks and operations required.

As will be clear from Chapter 2, it is necessary to use discrete building blocks of geographical data that are capable of representing entities and continuous fields, as well as being able to represent exact or inexact attributes, location, and relations. The main spatial units used in representing the data were shown to be the vector (point, line, and polygon) and raster (pixel) primitives; the different data types associated with each kind of data model were also highlighted. For organizing their storage within the computer, these basic building blocks may be formalized into logical schemata (known as data structures), which are used for representing both the entity and continuous data models. Table 3.1 summarizes the properties of the different data structures with respect to entity locational data, attribute storage, and topology. Information formulated according to one geographical data model does not necessarily have to be organized in the computer using a data structure of the same name. Raster data structures may be used for representing data formulated according to the vector data model, and vice versa.

Table 3.1. Characteristics of basic spatial entities in vector and raster mode

Basic unit type	Locational data	Attributes held as	Topology
Vector units			
Points	*X,Y,(Z)*	records	implicit
Nodes	*X,Y,(Z)*	records	explicit
Simple line	*N*[*X,Y,Z*]	records	implicit
Complex line (arc)	*N*[*X,Y,Z*]	*N*[records]	explicit
Polyhedron (solid body)	*M*[lines]	records	explicit
Raster units			
Grid cell (pixel)	row, column	single value	implicit
Line	*N*[pixels]	single values or pointer to records	implicit
Polygon	*M*[pixels]	single values or pointer to records	implicit
Voxel	row, column, layer	single values or pointer to records	implicit

POINTS, LINES, AND AREAS: VECTOR DATA STRUCTURES

A data structure that uses points, lines, or polygons to describe geographical phenomena is known as a *vector data structure*. Vector units are characterized by the fact that their geographical location may be independently and very precisely defined, as may be their topological relationships. They usually admit no internal variation (unless they are composed of simpler units) so that their attributes refer to the whole unit.

A spatial phenomenon is modelled in terms of its geographic location and attributes. An oil well could be represented by a point unit consisting of a single XY coordinate pair and the label 'oil well'; a section of oil pipeline could be represented by a line unit consisting of a starting *XY* coordinate and an end *XY* coordinate and the label 'oil_pipeline'; an oil refinery could be represented by a polygon unit covering a set of *XY* coordinates plus the label 'oil_refinery'. The labels could be the actual names as given here, or they could be numbers that cross-reference with a legend or a table of additional information, or they could be special symbols.

Simple spatial units are rarely sufficient for efficient data handling. It is very useful to be able to handle more complex phenomena as though they were single units in the information system. The simplest complex object is the 'string', 'chain', or '*arc*' (GIS terminology is rich in synonyms and near-synonyms) that is used to represent a curved or non-straight line. When lines join, as at a river junction, or road intersection, they must be linked by a special kind of point called a *node*, whose function is to carry information about the way the lines are linked together and how traffic could flow from one to the other. If lines cross without being joined by nodes the computer cannot distinguish between a bridge, an underpass, or a crossroads.

Nodes are also used to carry information about how the boundary arcs of polygons link up and provide information about left and right polygon neighbours and enclosed islands. The methods used to link arcs into polygon networks are described in more detail later in this chapter. Attributes of vector units are stored in computer files as *records* or *tuples* that may be linked to them by pointers.

Typical operations on the geometric data combine layers of vector information to derive answers to Boolean type questions or to quantify relations (geometric or topological) between various units within and across layers. Attribute-based operations allow searching the occurrence of a particular set of values, or for new values to be computed from existing data.

GRID CELL: RASTER DATA STRUCTURES

Spatial phenomena may also be represented by sets of regular shaped tessellated units as described in Chapter 2. The simplest form is the square cell (pixel) and the tessellated regular grid is known as a *raster data structure*. The location of entities is defined with direct reference to the grid network with each pixel associated with a square parcel of land on the earth's surface. The resolution or scale of the raster data is then

Table 3.2. Data structures and computer coding

Data structure	Computer coding of data types
Vector units	
geometric data	x,y,z coordinates are usually scalar or real data types using 32 or 64 bits
attribute data	are highly variable and may be binary, nominal, ordinal, integer or real, or directional using 16, 32, or 64 bits
topological data	usually small numbers so 16 bits are often enough
Raster units	
grid locations and size	32-bit real numbers
attributes	16-bit integers
	32- or 64-bit reals for mathematical modelling

the relation between the pixel size and that of the cell on the ground. Whereas in vector data structures the topology between different units is explicitly recorded through database pointers (explained later in this chapter), in raster databases this is only implicitly coded through the attribute values in the pixels.

The variation of the phenomena may be represented in the grid array with a different real-valued number per pixel; with each attribute being described by a separate overlay. In more complex raster structures, cell values could be referenced by a pointer to a record carrying the value of many separate attributes. New attributes may be created from existing attributes (e.g. by classification) or by carrying out various kinds of spatial search or equivalent operation. Measured data are rarely available to generate these surfaces directly so interpolation techniques are often used to convert sampled points into continuous surfaces (see Chapter 5).

Typical operations include Boolean retrieval and neighbourhood querying. In addition mathematical modelling is relatively easy to undertake with the raster data structure by combining attribute data from various layers for the same or neighbouring raster cells.

Computer storage of vector and raster data structures

The storage requirements of these different data types are summarized in Table 3.2.

Database structures: data organization in the computer

Before considering in detail the ways in which spatial entities may be stored efficiently in the computer, we must first consider the general issues of organizing data for optimal storage and access. Although it is not essential for users of GIS to understand in detail how the data may be ordered inside the computer, any more than the driver of a car needs to know about the workings of the internal combustion engine, a little knowledge of data modelling and data structuring methods will help them to understand better how the systems work, and what their limitations and advantages might be. The following section presents only a brief introduction. Interested readers should consult a standard work in information storage and retrieval such as Date (1995), Salton and McGill (1983), Ullman (1980), Whittington (1988), or Wirth (1976).

File and data access

The essential features of any data storage system are that they should allow data to be accessed and cross-referenced quickly. There are several ways of achieving this, some of which are more efficient than others. Unfortunately, there seems to be no single 'best' method that can be used for all situations, which explains in part the massive investment in manpower and money in effective database management systems, as the computer programs are known that control data input, output, storage, and retrieval from a digital database.

Simple lists

The simplest form of database is a simple list of all the items. As each new item is added to the database it is simply placed at the end of the file, which gets longer and longer. It is very easy to add data to such a system, but retrieving data is inefficient. For a list containing n items, it takes an average of $(n + 1)/2$ search operations to find the item you want. So, for an information system containing 10 000 descriptions of items on cards, given that it takes 1 second to read the card name or number, it would take an average of (10 000 + 1)/2 seconds or about an hour and a half to find the card you want.

Most people know that searching through unstructured lists trying to find 'the needle in the haystack' is very inefficient. It is thus an obvious step to order or structure the data and to provide a key to that structure in order to speed up data retrieval.

Ordered sequential files

Words in a dictionary or names in a telephone book are structured alphabetically. Addition of a new item means that extra room must be created to insert it, but the advantages are that stored items may be reached very much faster. Very often ordered sequential files are accessed by binary search procedures. Instead of beginning the search at the beginning of the list, the record in the middle is examined first. If the key value (e.g. the sequence of letters in a word) matches then the middle record is the one being sought. If the values do not match, a simple test is made to see whether the required item occurs before or after the middle element. The appropriate half of the file is retained and the search repeated, until the item has been located. Binary search requires $\log2(n + 1)$ steps. If the file is 10 000 items long, and the search time per item is 1 second, the average time to find an item is approximately 14 seconds, compared with the 1.5 hours previously!

Indexed files

Simple sequential and ordered sequential files require that data be retrieved according to a key attribute. In the case of a dictionary, the key attribute is the spelling. But in many applications, particularly in geographical information, the individual items (pixels, points, lines, or areas) will have not only a key attribute such as an identification number or a name, but will also carry information about associated attributes. Very often it is the information about the associated attributes that is required, not just the name. For example, we may have an ordered list of soil profiles that has been structured by soil series name, but we would like to retrieve information about soil depth, drainage, pH, texture, or erosion. Unless we adopt another database strategy, our search procedures revert to those of the simple sequential list.

With indexed files, access to the original data file can be speeded up in two ways. If the data items in the files themselves provide the main order of the file, then the filers are known as *direct files* (see Table 3.3*a*). The location of items in the main file may also be specified according to topic, which is given in a second file, known as an *inverted file* (see Table 3.3*b*). Just as in a book, examination of the index can determine the items (pages) which satisfy the search request.

In the direct file, the record for each item contains sufficient information for the search to jump over unnecessary items. For example, consider a data file containing soil series names that have been ordered alphabetically. Each item contains not only the series name and other information but also a number indicating the storage location of series names beginning with that same letter. The search for a particular record is then made much simpler by constructing a simple index file that lists the correspondence between first letter of series name and storage location. The search proceeds by a sequential search of the index, followed by a sequential search of the appropriate data block. The average number of search steps is then $(n_1 + 1)/2 + (n_2 + 1)/2$, where n_1 is the number of

Table 3.3.

(*a*) Indexed files

Index		File item
Item key	Record Number	
A	1	A_1
B	$n_a + 1$	A_2
C	$n_a + n_b + 1$	...
...	...	B_1
...	...	...
...	...	C_1

(*b*) Inverted files

Soil profile number	Attributes					
	S	pH	De	Dr	T	E
1	A	4	deep	good	sandy	no
2	B	5	shallow	good	clay	yes
3	C	6	shallow	poor	sandy	no
4	D	7	deep	good	clay	yes
5	E	4	deep	poor	clay	no
6	F	5	shallow	poor	clay	no

S = soil series, De = depth, Dr = drainage, T = texture, E = erosion.

(*c*) Index inverted file

Topic	Profiles (sequential numbers in original file)					
Deep	1			4	5	
Shallow		2	3			6
Good drainage	1	2		4		
Poor drainage			3		5	6
Sandy	1		3			
Clay		2		4	5	6
Eroded		2		4		

steps in the index and n_2 is the number of items in the data block referenced by the index.

The use of the inverted file index requires first that it be constructed by performing an initial sequential search on the data for each topic. The results are then assembled in the inverted file or index, which provides the key for further data access (see Table 3.3*c*).

Indexed files permit rapid access to databases. Unfortunately they have inherent problems when used with files in which records are continually being added or deleted, such as often happens with interactive mapping systems. Addition or deletion of a record in a direct file means that both the file and its index must be modified. When new records are written to a file accessed by an inverted file index, the new record does not have to be placed at a special position; it may be simply added to the end of the file, but the index must be updated. File modification may be an expensive undertaking with large data files, particularly in an interactive environment.

A further disadvantage of indexed files is that very often data can only be accessed via the key contained in the index files; other information may only be retrievable using sequential search methods.

Database structures and database management

As the above examples show, data on real world entities are written into computer files, which are essentially simple or organized lists. The basic unit of a file is the *data record* or *tuple* which contains all the relevant information for each entity. Depending on the kinds of data collected, the data records may all be of the same length, or may be of variable length, and measures are taken to ensure that both types can be efficiently encoded (Box 3.2). Spatial databases, however, contain many files with data on related aspects of the same entities, or of data on entities that because of their spatial proximity or connectivity have to be linked or grouped together. It is essential to organize the way these files are stored and linked in the computer to model real world phenomena and to ensure efficient storage and retrieval of data. A computer program that is designed to store and manage large amounts of data is called a *database management system (DBMS)*.

Modern DBMS use many methods for efficiently storing and retrieving data (e.g. Date 1995, Worboys 1995) but all are based on three fundamental ways of organizing information that also reflect the logical models used to model real world structures: these are known as the *hierarchical, network*, and *relational schemata*, and they are all used in GIS. Recently a fourth structure has been used for CAD-CAM and GIS, which is called Object-Orientation (referred to by some as 'O-O' for short). O-O is a further development of the network model which enables the interconnectivity and ownership relations between spatial entities to be modelled effectively. Object orientation is now used in some commercial GIS, although research on developing the concept is still ongoing.

Database modelling is the task of designing a database that performs efficiently, contains correct information, has a logical structure and is as easy as possible to maintain and extend (Worboys *et al.* 1990). In order to understand how this can be achieved in GIS it is first necessary to examine the fundamental principles behind the different organizational structures and to see how they can be used for spatial information that has been recorded using either the exact entity or the continuous field data models.

The hierarchical database structure

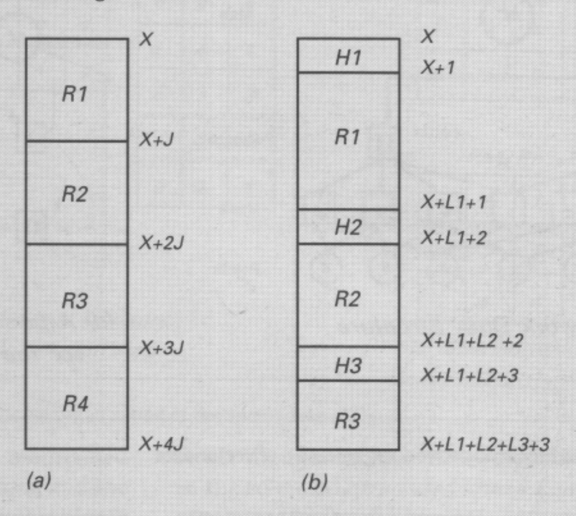
When the data have a parent/child or one-to-many relation, such as areas for various levels of administration, soil series within a soil family, or pixels within a polygon, hierarchical methods provide quick and convenient means of data access. Hierarchical systems of data organization are well known to the environmental sciences as they are the methods used for plant and animal taxonomies, soil classifications, and so on. They assume that each part of the hierarchy can be reached using a key (a set of discriminating criteria) that fully describes the data structure. The assumption is also made that there is a good correlation between the key attributes (discriminating criteria) and the associated attributes that the items may possess.

Figure 3.1*a* presents a simple example of a 2-polygon map and Figure 3.1*b* shows how it could be coded according to a hierarchical model. The main

BOX 3.2. ADDING INFORMATION ABOUT DATA LOCATION IN A HEADER RECORD IMPROVES EFFICIENCY OF DATA STORAGE. (*A*) NO HEADERS—ALL DATA RECORDS ARE OF EQUAL LENGTH WHETHER FULL OR NOT, (*B*) WITH HEADERS RECORD LENGTHS MATCH THE AMOUNT OF DATA

Data Records

In all kinds of database structures, the data are written in the form of records. The simplest kind of record is a one-dimensional array of fixed length, divided into a number of equal partitions, as shown in (*a*). This record is ideal when all items have the same number of attributes, for example when a number of soil profiles have been sampled and analysed for a standard range of cations.



Fixed length records are inconvenient, however, when the attributes are of variable length, and when the set of attributes measured is not common to all items. For example, not all soil profiles have the same number of horizons, and not all polygon boundaries have the same number of coordinates. In these situations variable length records are used. Each record has a 'header', an extra attribute that contains information about the type of information in the sub-record and the amount of space it takes up, shown in (*b*). A series of records make up a file.

advantages of the model are its simplicity and ease of access via keys that define the hierarchy. They are easy to understand and to update and expand and can be useful for organizing data in mass storage systems (e.g. see Kleiner and Brassel 1986). Data access via the key attributes is easy and quick , but unfortunately is very difficult for associated attributes. Consequently, hierarchical systems are good for data retrieval if the structure of all possible queries is known beforehand. This is commonly the case with bibliographic, bank, or airline retrieval systems. For environmental data, however, the exploratory nature of many retrieval requests cannot be accommodated by the rigid hierarchy and the user rejects the system as impossibly inflexible. For example, Beyer (1984) reports that the Royal Botanical Gardens in Kew initially set up a hierarchical database for their herbarium collection of more than a million items based on the rules of plant taxonomy. This seemed sensible until the director wished to make a trip to Mexico to gather new material for the herbarium and asked the database which plants Kew already had from that land. Unfortunately for the suppliers and the users of the database, the director received no answer because the attribute

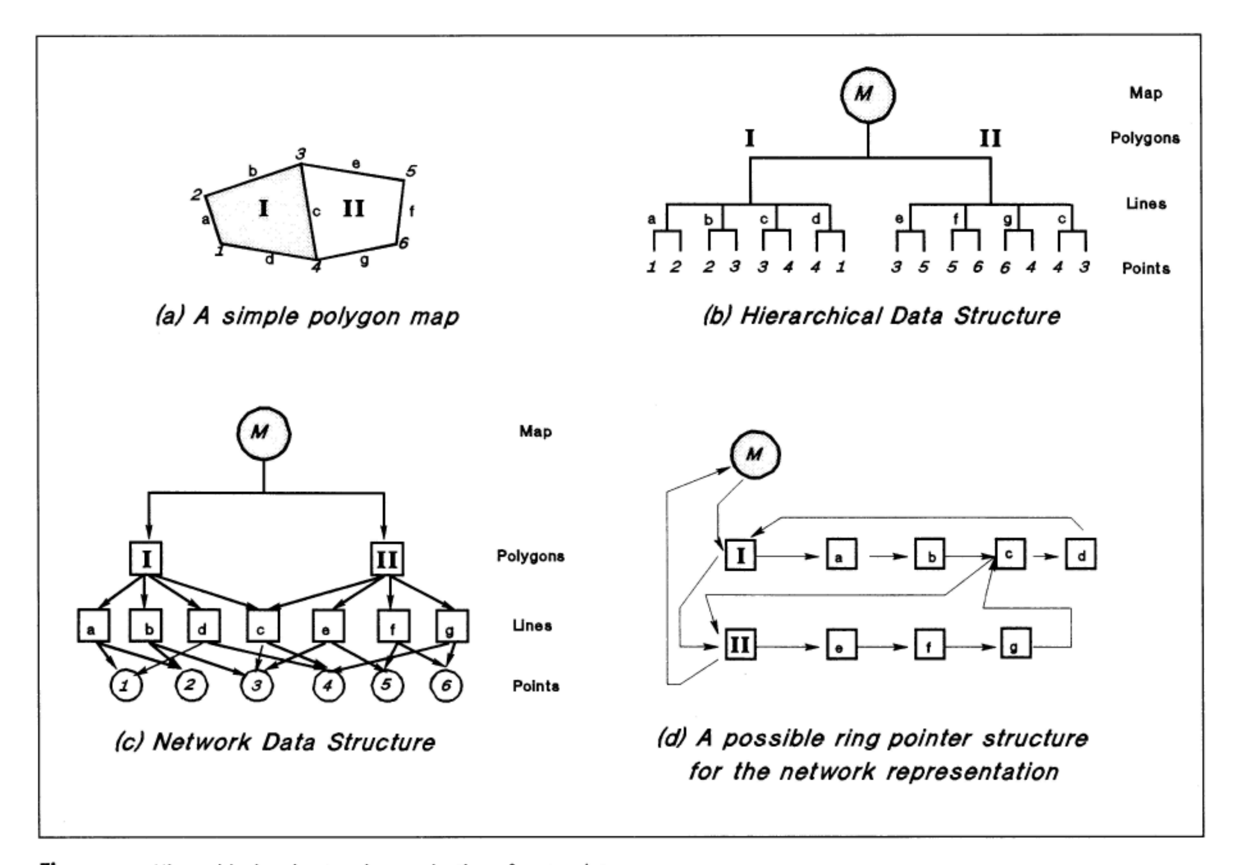


Figure 3.1. Hierarchical and network organization of vector data

'place of collection' is not part of the plant taxonomy key!

Further disadvantages of hierarchical database structures are that large index files have to be maintained, and certain attribute values may have to be repeated many times, leading to data redundancy, which increases storage and access costs. Figure 3.1*b* shows this redundancy as each coordinate pair would have to be repeated twice, and coordinates 3 and 4 would have to be repeated four times because line c has to be repeated twice. Hierarchical structures are also wasteful of space; for example, if an operation were made to give polygons I and II the same name there is no easy way to suppress the display of line c, which would become unnecessary for display.

THE NETWORK DATABASE STRUCTURE

In hierarchical structures, travel within the database is restricted to the paths up and down the taxonomic pathways. In many situations much more rapid linkage is required, particularly in data structures for graphics features where adjacent items in a map need to be linked together even though the actual data about their coordinates may be written in very different parts of the database.

Both redundancy and linkage problems are avoided by the compact network structure shown in Figure 3.1*c*, in which each line and each coordinate need appear only once. With this structure it is a simple matter to suppress the printing of line *c* whenever it is referenced by polygons having the same name, thus making map generalization easier. Very often in graphics, network structures are used that have a ring pointer structure. Ring pointer structures (Figure 3.1*d*) are very useful ways of navigating around complex topological structures such as polygon networks.

Network database structures are very useful when the relations or linkages can be specified beforehand. They avoid data redundancy and make good use of available data. The disadvantages are that the database is enlarged by the overhead of the pointers, which in complex systems can become quite a substantial part

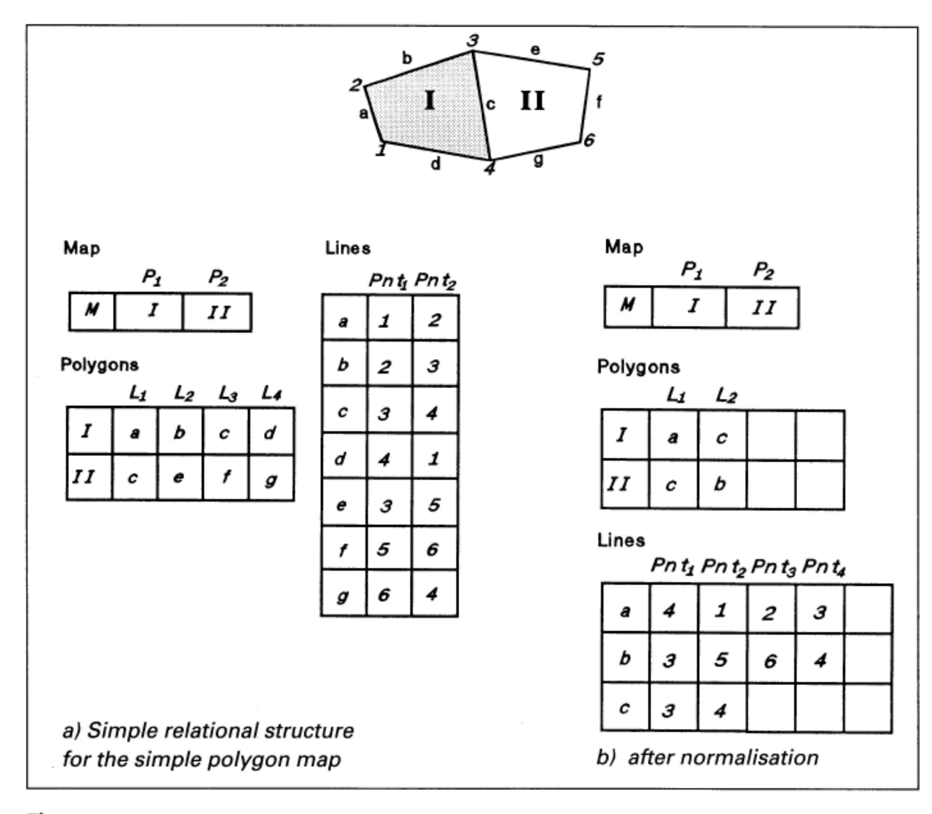


Figure 3.2. Relational organization of the vector data of Figure 3.1

of the database. These pointers must be maintained every time a change is made to the database and the building and maintenance of pointer structures may be a considerable overhead for the database system.

THE RELATIONAL DATABASE STRUCTURE

In its simplest form the relational database structure stores no pointers and has no hierarchy. Instead, the data are stored in simple records, known as *tuples*, which are sets of *fields* each containing an attribute; tuples are grouped together in two-dimensional tables, known as *relations*, somewhat in the manner of a spreadsheet program. Each table or relation is usually a separate file. The pointer structures in network structures and the keys in hierarchical structures are replaced by data redundancy in the form of identification codes that are used as unique keys to identify the records in each file (Figure 3.2*a*).

Data are extracted from a relational database through the user defining the relation that is appropriate for the query. This relation is not necessarily already present in the existing files, so the controlling program uses the methods of relational algebra to construct the new tables. These rules are often encoded in the so-called *Structured Query Language (SQL)*—see Date (1995) and Chapter 6. Figure 3.2*a* shows that the relational structure of the simple map of Figure 3.1*a* also contains much redundancy and methods known as *normalization* are used to create more efficient codings. For example, addressing the lines as sets of straight lines segments (*arcs*) simplifies the relational structure without loss of information (Figure 3.2*b*).

Relational databases have the great advantage that their structure is very flexible and may meet the demands of all queries that can be formulated using the rules of Boolean logic and of mathematical operations. They allow different kinds of data to be searched, combined, and compared. Addition or removal of data is easy too, because this just involves adding or removing a tuple, or even a whole table (as shown in Figure 3.3). Querying across different relational tables is made by joining them through common fields. This is good for situations where all records have the same number attributes and there is no natural hierarchy. However, where the relationships between tables are complex and a number of joins are needed, operations take a

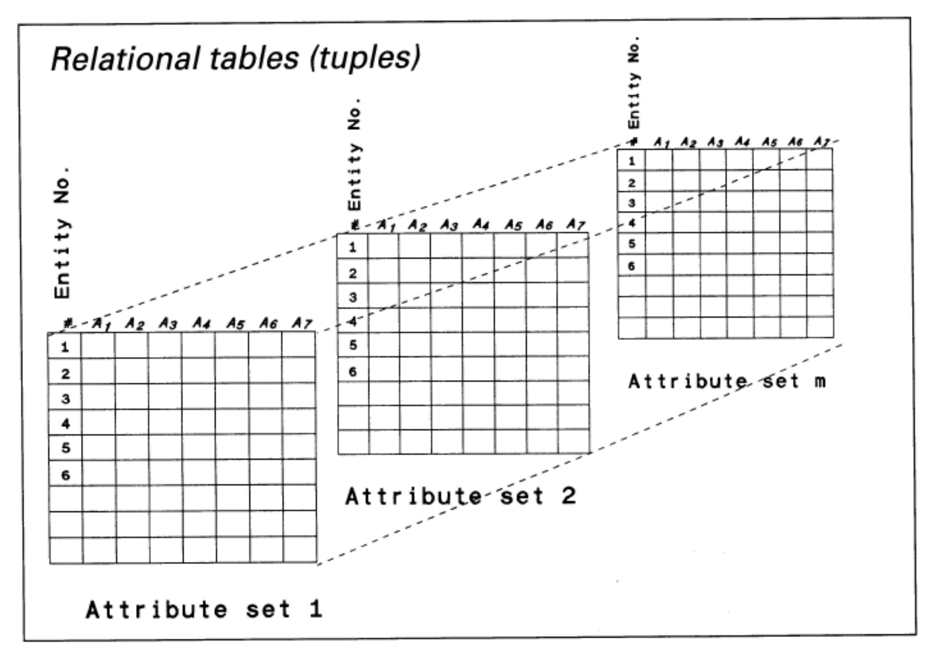


Figure 3.3. Adding new data to a relational system is just a matter of defining new tables

considerable amount of computer time even on fast computers. Consequently, relational database systems have to be very skilfully designed in order to support the search capabilities with reasonable speed, which is why they are expensive. They were first used for GIS in the early 1980s (Abel 1983, Lorie and Meier 1984, van Roessel and Fosnight 1984) and they are now established as major tools for handling the attributes of spatial entities in many well-known commercial GIS (see Hybrid databases, below).

THE OBJECT-ORIENTED DATABASE STRUCTURE

Object-oriented concepts originated in programming languages such as Simula (Dahl and Nygaard 1966) and Smalltalk (Goldberg and Robson 1983) and the application of these ideas to databases was stimulated by the problems of redundancy and sequential search in the relational structure. In GIS their use has been stimulated by the need to handle complex spatial entities more intelligently than as simple point, line, polygon primitives, and also by the problems of database modifications when analysis operations like polygon overlay are carried out (see Chapter 7). The approach is being applied increasingly in a number of fields although there are many different formalisms of the concept with no clear agreement on the definition amongst the computing community (Worboys *et al.* 1990). The terms of 'object-based' or 'object-centred' are more nebulous in meaning than the definable characteristics of 'object-oriented programming languages' (Jacobsen *et al.* 1992).

Object-oriented database structures, developed using object-oriented programming languages, combine the speed of hierarchical and network approaches with the flexibility of relational ones by organizing the data around the actual entities as opposed to the functions being processed (Chance *et al.* 1995, Kim and Lochovsky 1989). In the relational structure, each entity is defined in terms of its data records and the logical relations that can be elucidated between the attributes and their values. In object-oriented databases, data are defined in terms of a series of unique objects which are organized into groups of similar phenomena (known as object classes) according to any natural structuring. Relationships between different objects and different classes are established through explicit links.

The characteristics of an object may be described in the database in terms of its attributes (called its state) as well as a set of procedures which describe its behaviour (called operations or methods). These data are encapsulated within an object which is defined by a unique identifier within the database. This remains the same whatever the changes are to the values that describe their characteristics (Bhalla 1991). For

example, a building 'object' might change over time in terms of structure (or use) but its unique identifier will remain the same. The relational model of a bicycle is merely a list of parts; the O-O model links the parts so that their function in relation to each other and the behaviour of the object is clearly expressed.

The structuring of objects within the database is established using pointers which refer directly to the unique identifiers. The classes and instances within them are linked by the pointers to show various relationships and hierarchies. Where hierarchies are established, forming general, sub, and super classes, various defined states or methods are passed down through the system of inheritance. This means that efficiencies may be made both in characterizing the attributes of objects and in retrieving them from the database. Figure 3.4 compares a relational database for a utility company with an object-oriented approach. In the relational approach each main table (District, User #, Address) is linked by data that are repeated from one table to the next. In the object-oriented approach, 'District', 'Address' and 'User #' are designated as 'objects', with defined relations between them, such as 'contains', 'is situated in', 'used by', 'lives at', etc. All data are held once only, and the directed pointers serve not only for rapid retrieval of data, but also for passing commands up and down the database hierarchy. For example, it is much easier with the O-O approach to find all users of a given type in a given district in order to adjust their billing arrangements.

Once the data have been encapsulated as an object in the database the only way to change them or to query them is to send a request, known as a message, to carry out one of its operations (Chance *et al.* 1995). The types of querying possible depends on the operations that have been used to define the objects. The response of the object to a message will depend on its state and the same message may bring about a different reaction when received by different objects or in a different context; this is termed polymorphism.

Data used in object-oriented databases need to be clearly definable as unique entities. Given that, these databases (as with their network and hierarchical counterparts) provide very efficient structures for organizing hierarchical, interrelated data. Establishing the database is obviously time-consuming as the objects may be defined more explicitly and the various links need to be established. Once this is finished, the database provides a very efficient structure for querying especially with reference to specific objects. That said, relational database structures are still better at

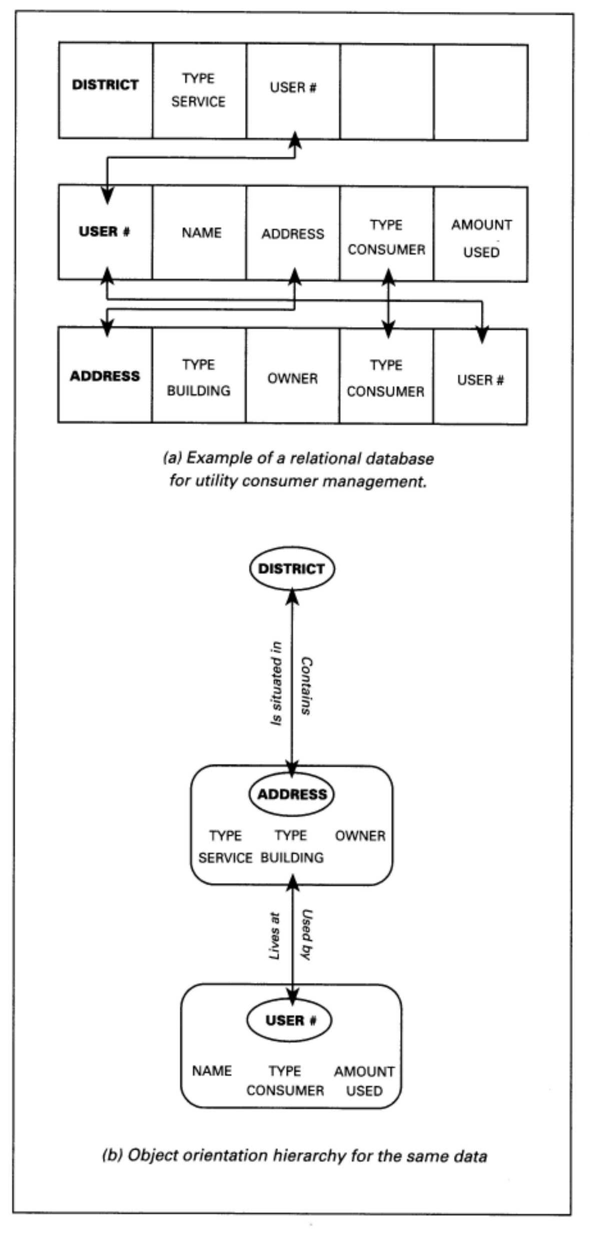


Figure 3.4. (*a*) Relational tables for a utility database include redundant information. (*b*) The object-oriented approach reduces data storage and improves links between the records

performing queries based on the values of an attribute although specialized indexing methods and clustering techniques are being developed for some (Arctur and Woodsford 1996).

The possibilities offered by object-oriented database structures for GIS have been explored by a number of researchers in the last few years (e.g. Worboys 1994)

and their use with entity-based data has been highlighted (Raper and Livingstone 1995). The application of this type of structuring is explored more fully later in this chapter.

New developments in database structures

Alternatives to the database structures mentioned above have been explored by various researchers seeking to represent more effectively and flexibly the spatial (and temporal) nature of geographical data. For example, Van Oosterom (1993) has described a nested approach whereby different scales of representation of data are stored in a database in which links are made between the same geographically referenced areas.

One active area of research has been in the use of the deductive database approach. This extends a logic-based approach to information storage, querying, and processing using programming languages such as Prolog. A deductive database stores both the data and the logic that defines a fact or which expresses a relationship; data records are made up of an extension (which is similar in form to a relation database) and an intension part (which is a virtual relation expressed logically using other relations) (Yearsley and Worboys 1995, Worboys 1995, Quiroga *et al.* 1996). The advantage of the deductive database approach is that it allows more complex objects and modelling to be constructed than may be undertaken using relational, network, or hierarchical structures and so offer great potential for spatial data handling.

The development of GIS based on these database structures is still essentially at the theoretical stage at present.

Database management systems

Database management systems (DBMS) are computer programs for organizing and managing the database and they may be constructed using any of, or a combination of, the hierarchical, network, relational, and object-oriented structures presented above. The aim of the database management system is to make data quickly available to a multitude of users whilst still maintaining its integrity, to protect the data against deletion and corruption, and to facilitate the addition, removal, and updating of data as necessary. According to Frank (1988), a database management system should provide the following functionality:

(*a*) Allow storage and retrieval of data and data selection based on one or more attributes or relations.

(*b*) Standardize access to data, and separate data storage and retrieval from the use of data in application programs to maintain independence in those programs.

(*c*) Provide an interface between database and application program based on a logical description of the data without requiring details of the physical storage.

(*d*) Make access functions in applications independent of the physical storage structure so that programs are not affected by changes in storage media.

(*e*) Allow several users to access the data simultaneously.

(*f*) Protect the database from indiscriminate and illegal changes.

(*g*) Provide sound rules for data consistency which will be enforced automatically. These rules are an excellent way of removing errors and inconsistencies from the database.

Most database management systems allow access to data through a high-level programming language and through a user-friendly query language, of which SQL (Structured Query Language) is the most common for large relational database management systems. The user interface allows casual interrogation of the database while the high-level programming interface allows the database to be linked directly to application programs such as GIS.

A good DBMS will also ensure that spatially contiguous data will be stored in physically adjacent areas of the storage medium in order to reduce the data access and transfer times of the computer. Speedy access to large amounts of data can be a critical aspect of designing GIS that operate smoothly in response to users' demands. Many GIS use the DBMS as part of their system to take advantage of these data handling capabilities.

Choosing the most appropriate database structure

It should be apparent that the four basic database structures—hierarchical, network, object-oriented, and relational—all have something to offer for spatial information systems. Hierarchical systems allow large databases to be divided up easily into manageable chunks, but they are inflexible for building new search paths and they may contain much redundant data. Network systems contain little redundant data, if any, and provide fast, directed, if inflexible links between related entities. Object-oriented systems permit relations, functionality, persistence, and interdependence to be built into one system at the expense of the programming tools being more complex and heavier demands on computing power. Relational systems are open, flexible, and adaptable, but may suffer from large data volumes, redundancy, and long search times. Consequently it is not surprising that these techniques are often used together in spatial information systems to complement each other, rather than assigning all the work to one of them.

A hierarchical approach is often useful for dividing spatial data into manageable themes, or into manageable areas, so that continuous, seamless mapping becomes possible. The network approach is ideal for topologically linked vector lines and polygons. The relational approach is good for retrieving objects on the basis of their attributes, or for creating new attributes and attribute values from existing data. Object orientation is useful when entities share attributes or interact in special ways.

Data structures for representing the geometry of spatial phenomena

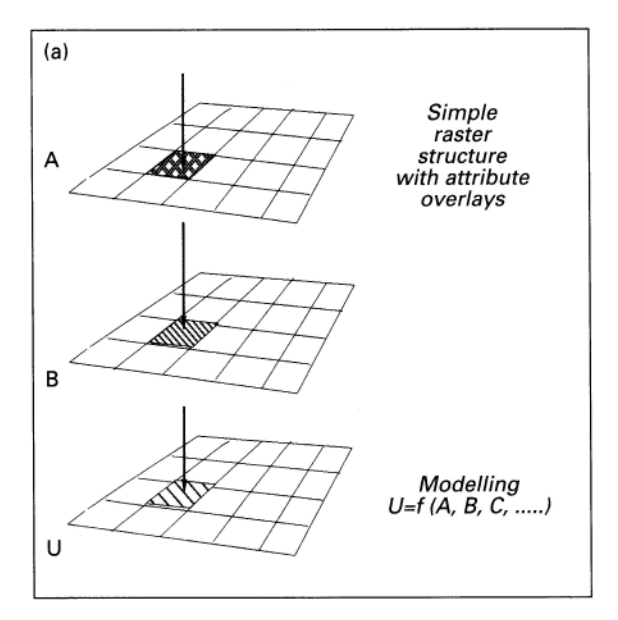
These main database structures influence how geographical data are discretized and stored in a computer system such as a GIS. Many systems, either raster or vector, make use of the 'overlays' or 'feature planes' methods of structuring spatial data into intuitively useful groups.

Data organization in raster data structures

A raster database is built up from what the user perceives as a number of Cartesian overlays; given the large number of coordinates organizing schemata aim to optimize data access and to minimize storage and processing requirements (as shown in Figure 3.5*a*).

In simple raster structures where each cell on each overlay is assumed to be an independent unit in the data base (one-to-one relation between data value, pixel, and location), each cell is identified by a coordinate pair and a set of attribute values for each overlay as shown in Figure 3.6*a*. This is a very data hungry structure with no data held on cell size, or display symbols, and no compression techniques used to reduce storage demands. In an alternative method (Figure 3.6*b*) each overlay may be represented in the database as a two-dimensional matrix of points carrying the value of a single attribute. This still requires much storage space as it contains lists of redundant coordinates which are repeated for every overlay and again no data on cell size or display symbols are held.

The hierarchical structure shown in Figure 3.6*c* (used in the Map Analysis Package of Tomlin 1983) establishes a many-to-one relation between attribute values and the set of points in the mapping unit so uniform areas (polygons) may be addressed easily. Recoding or changing variables such as the display symbols is made easy as it requires rewriting only one number per mapping unit per overlay as opposed to all cell values with the previous two structures. This structuring also allows run length code and quadtree data compression techniques (discussed later in this chapter) to be used to reduce storage demands and is efficient in handling structured data. The SPANS GIS is based on a quadtree form of this hierarchical structure. The main disadvantage of this structure is that it is clumsy for highly variable field data.



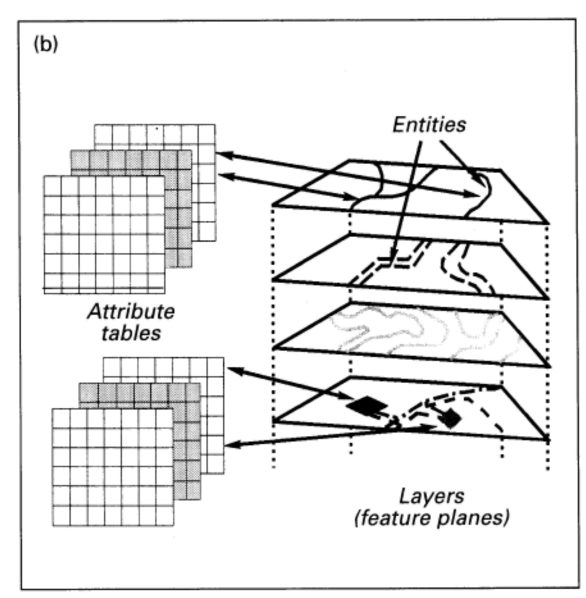


Figure 3.5. Data planes and feature layers in (*a*) raster and (*b*) vector data structures

With the fourth structure each overlay is stored as a separate file with a general header containing information such as map projection, cell size, numbers of rows and columns, and data type; this is followed by a simple list of values which are ordered according to the sequence of rows and columns (Figure 3.6*d*). This is obviously more efficient as coordinate values are not stored for each cell and generic geometry and display values are written in the header of the overlay. PCRaster (Wesseling *et al.* 1996) uses this structure.

Compact methods for storing raster data

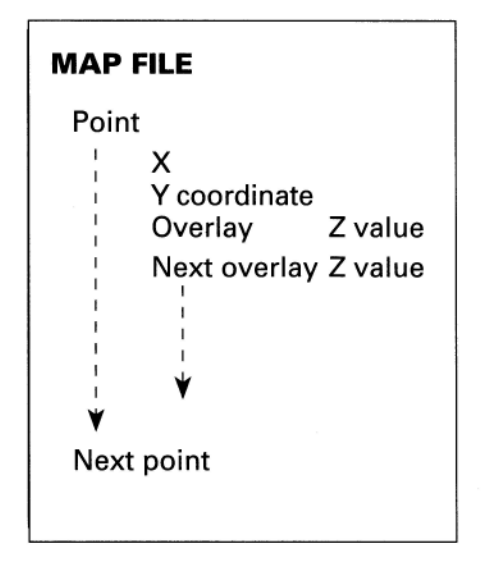
When raster data structures are used to represent a continuous surface, each cell has a unique value and it takes a total of *n*rows × *m*columns to encode each overlay, plus general information on map projection, grid origin, grid size, and data type. The most storage-hungry data layer is one where the data type is a scalar and each cell contains a real number, as in the case of altitude matrices for digital elevation and other continuous surfaces obtained from interpolation. Other data types will require less space because they can code their data in fewer bits (as discussed earlier).

When raster structures are used to represent lines or areas in which the pixels everywhere have the same value, it is possible to effect considerable savings in the storage requirements for the raster data, providing of course that the data structures are properly designed. The structures described in Figure 3.6*a* and *b* may use array coordinates to reduce the actual quantity of numbers stored, with the limitation that all spatial operations must be carried out in terms of array row and column numbers. These systems do not encode the data in the form of a one-to-many relation between mapping unit value and cell coordinates, so compact methods for encoding cannot be used.

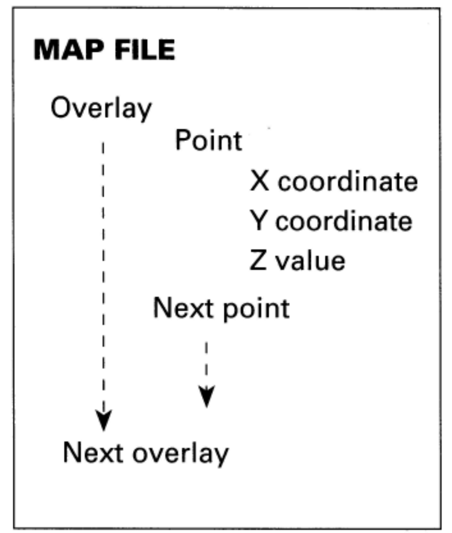
The third structure given in Figure 3.6*c* references the sets of points per region (or mapping unit) and allows a variety of methods of compact storage to be used. There are four main ways in which the spatial data for a mapping unit or polygon may be stored more economically: these are chain codes, run-length codes, block codes, and quadtrees.

Chain codes Consider Figure 3.7. The boundary of region A may be given in terms of its origin and a sequence of unit vectors in the cardinal directions. These directions can be numbered (East = 0, North = 1, West = 2, South = 3). For example, if we start at cell row = 10, column = 1, the boundary of the region is coded clockwise by:

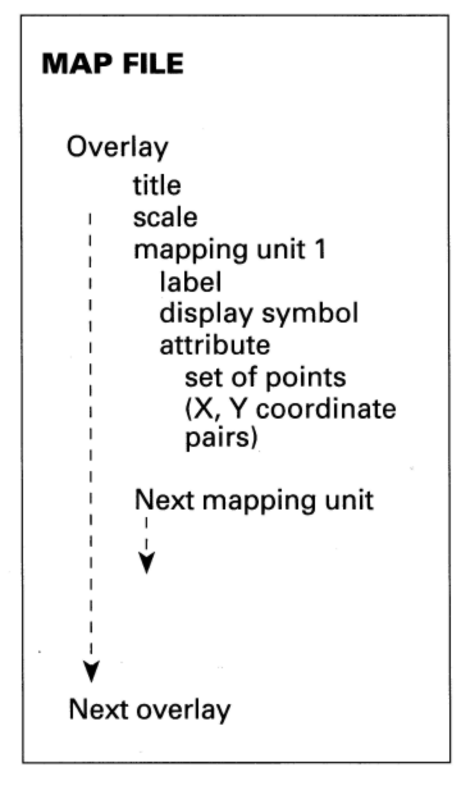
$$0, 1, 0^2, 3, 0^2, 1, 0, 3, 0, 1, 0^3, 3^2, 2, 3^3, 0^2, 1, 0^5, 3^2, 2^2, 3, 2^3, 3, 2^3, 1, 2^2, 1, 2^2, 1, 2^2, 1, 2^2, 1^3$$



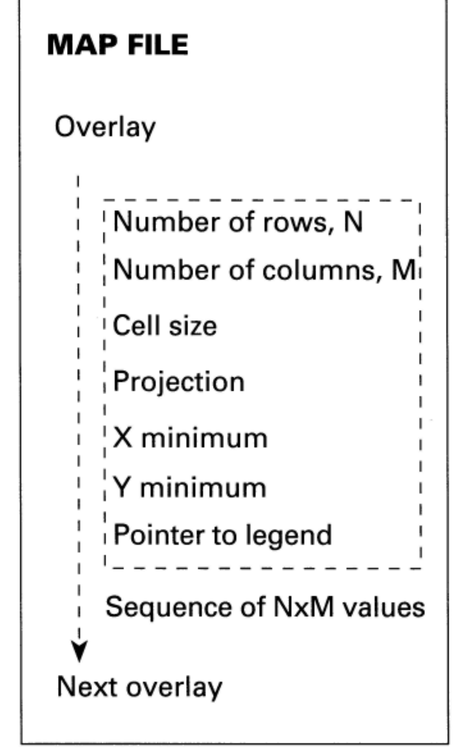
(a)



(b)



(c)



(d)

Figure 3.6. Four different ways of creating a raster data structure
(*a*) Each cell is referenced directly; (*b*) each overlay is referenced directly; (*c*) each mapping unit is referenced directly; (*d*) each overlay is a separate file with a general header

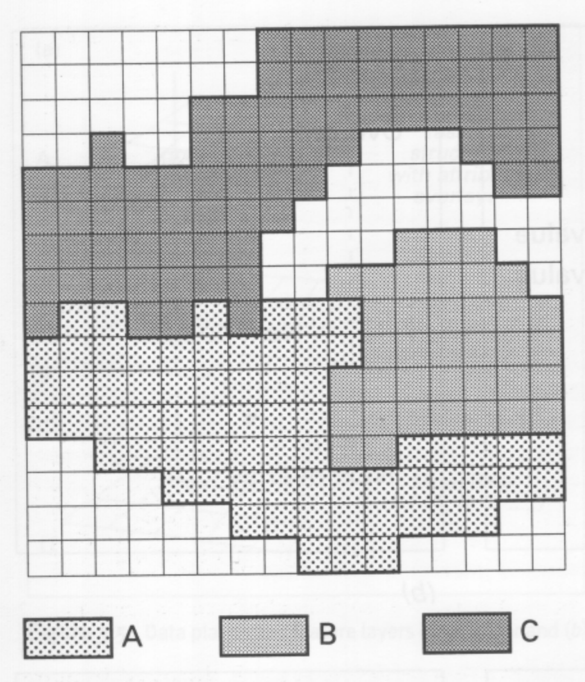


Figure 3.7. A simple raster map

where the number of steps (pixels) in each direction is given by the superscript number.

Chain codes can be stored using integer data types and therefore provide a very compact way of storing a region representation; they allow certain operations such as estimation of areas and perimeters, or detection of sharp turns and concavities to be carried out easily. They are also useful for converting the raster description of polygons to vector form, though the jagged steps must usually be smoothed away, thereby introducing errors in the location of the boundaries. Overlay operations such as union and intersection are difficult to perform with chain codes without returning to a full grid representation. Another disadvantage is the redundancy introduced because all boundaries between regions must be stored twice. Freeman (1974) gives further details.

Run length codes Run length codes allow the points in each mapping unit to be stored per row from left to right per row for each class encountered in terms of a begin cell, an end cell, and an attribute. The integer data type is usually all that is required, and even smaller integer representations at 8-bit word level could suffice for some applications.

For the region A shown in Figure 3.7 the codes would be as follows:

Row 9	2,3 6,6 8,10
Row 10	1,10
Row 11	1,9
Row 12	1,9
Row 13	3,9 12,16
Row 14	5,16
Row 15	7,14
Row 16	9,11 to be stored per row from left to right for each class.

In this example, the 69 cells of region **A** have been completely coded by 22 numbers, thereby effecting a considerable reduction in the space needed to store the data.

Clearly run-length coding is a considerable improvement in storage requirements over conventional methods whenever the many-to-one relations are present. It is especially suitable for use in situations where total volumes of data must be kept limited. On the other hand, too much data compression may lead to increased processing requirements during cartographic processing and manipulation. Run-length codes can also be useful in reducing the volume of data that need to be input to a simple raster database when large uniform areas need to be digitized (see Chapter 4). However, run-length codes are not suitable for coding continuous variation because each grid cell has a unique value and data compression is not possible or for variable levels of resolution.

Block codes The idea of run-length codes can be extended to two dimensions by using square blocks to tile the area to be mapped. Figure 3.8 shows how this can be done for region **A** of raster map of Figure 3.7. The data structure consists of just three numbers, the origin (the centre or bottom left) and radius of each square. This is called a *medial axis transformation* or MAT (Rosenfeld 1980). Region **A** may be stored by 17 unit squares + 9 4-squares + 1 16-square. Given that two coordinates are needed for each square the region may be stored using 57 numbers (54 for coordinates and 3 for cell sizes). Clearly, the larger the square that may be fitted in any given region and the simpler the boundary, the more efficient block coding becomes. Both run-length and block codes are clearly most efficient for large simple shapes and least so for small complicated areas that are only a few times larger than the basic cell. Medial axis transformation is used by fax machines to reduce the size of the images for transmission; the method also has advantages for performing union and intersection of regions and for detecting properties such as elongatedness (Rosenfeld 1980).

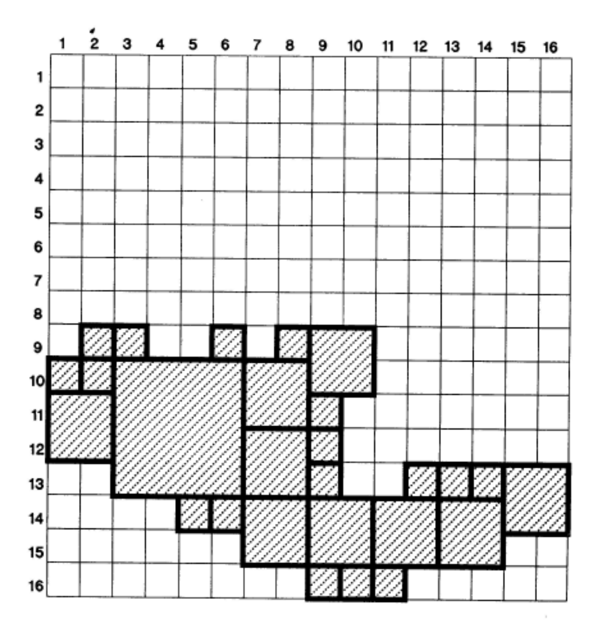


Figure 3.8. Medial axis transformation of polygon A in Figure 3.7

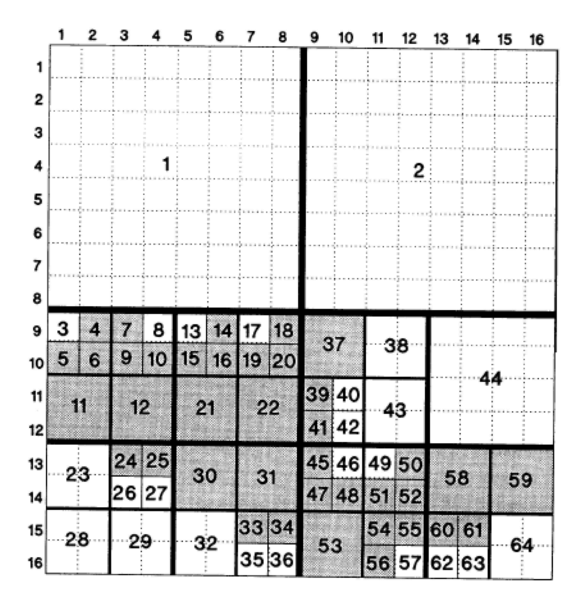


Figure 3.9. Quadtree encoding of polygon A in Figure 3.7

Quadtrees and binary trees One problem with regular grids is that the resolution of the data is limited by the size of the basic grid cell. Binary trees and quadtrees provide an approach to addressing successively finer levels of detail, with in principle, an infinite set of levels (Samet 1990*a*, 1990*b*).

The most efficient methods of compact representation of space are based on successive, hierarchical division of a $2^n \times 2^n$ array. If the division occurs by dividing the area into half each time, the method is known as a *binary tree* (and is discussed in more detail later); if the region is tiled by subdividing the array step by step into quadrants and noting which quadrants are wholly contained with the region the division is known as a *quadtree* and is the most used form. In both cases the lowest limit of division is the single pixel.

Figure 3.9 shows the successive division of region **A** (Figure 3.7) into quadrant blocks. This block structure may be described by a tree of degree 4, known as a *quadtree*; which is given in Figure 3.10. The entire

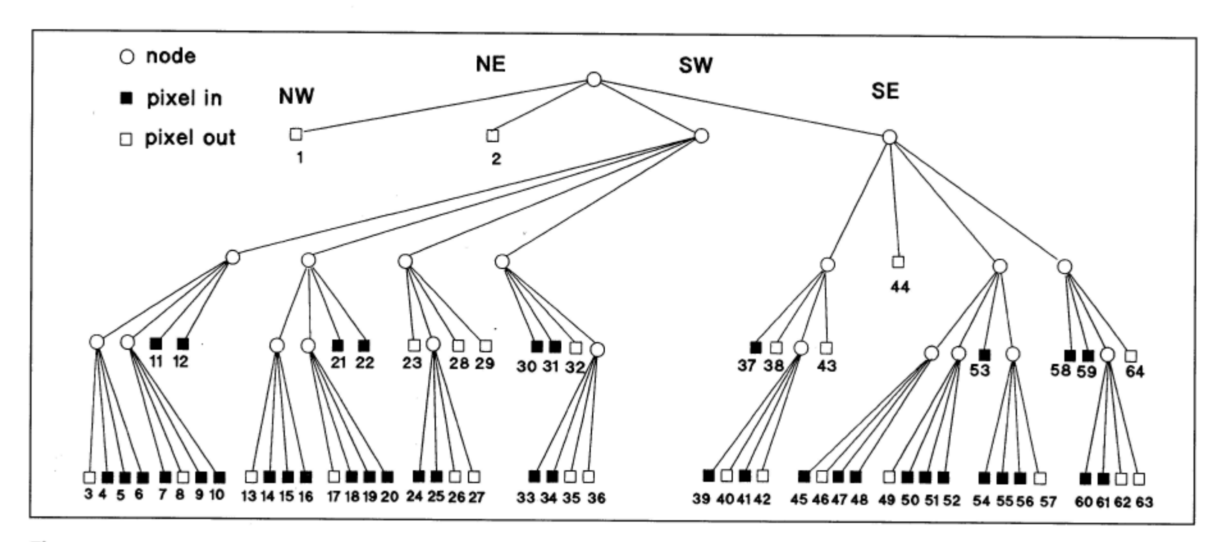


Figure 3.10. Quadtree hierarchy of polygon A in Figure 3.7

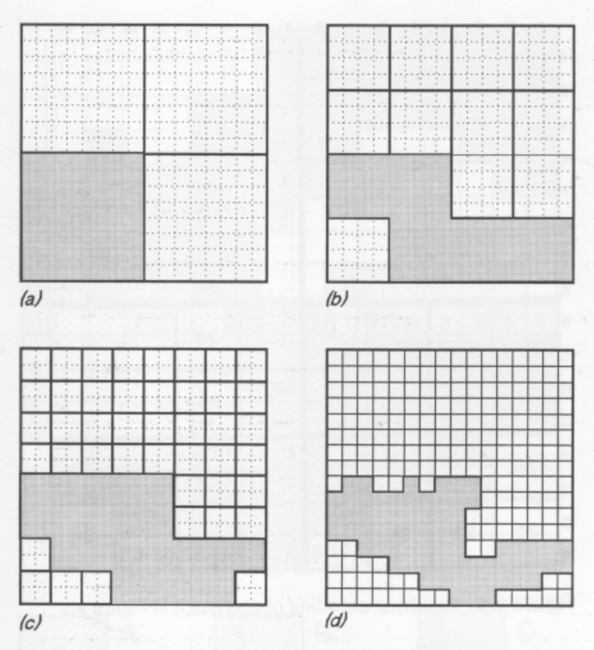


Figure 3.11. Quadtree hierarchies permit display at different levels of resolution

array of $2^n \times 2^n$ points starts from the root node of the tree, and the height of the tree is at most *n* levels. Each node has four sons, respectively the NW, NE, SW, and SE quadrants. Leaf nodes correspond to those quadrants for which no further subdivision is necessary.

During the 1980s there was much interest in the use of quadtrees in GIS (Martin 1982, Mark and Lauzon 1984) and it is clearly a technique that has much to offer. An authoritative description of the algorithms used for computing perimeters and areas, and for converting from raster-to-quadtree and other representations is given by Samet (1990*a*, 1990*b*). Quadtrees have many interesting advantages over other methods of raster representation. Standard region properties may be easily and efficiently computed. Quadtrees are 'variable resolution' arrays in which detail is represented only when available without requiring excessive storage for parts where detail is lacking (Figures 3.11*a–d*). In 3D systems, the *octtree* is the analogue of the 2D quadtree.

The largest problems associated with quadtrees appear to be that the tree representation is not translation-invariant—two regions of the same shape and size may have quite different quadtrees. Consequently shape analysis and pattern recognition are not straightforward, which is a problem for objects that move or change over time. Quadtree representation allows a region to be split up into parts, or to contain holes, without difficulty.

Two-dimensional orderings

Various ordering methods have been used with quadtrees and other pixel addressing systems, which reduce in effect the spatial referencing of the cells or defined areas down to a one-dimensional coordinate. This simplifies algorithms used in GIS and exploits various list data structures that are available so reducing demands on computer disk storage and memory (Abel and Mark 1990). The orderings define pathway directions through the two- or three-dimensional gridded space as shown in Figure 3.12; the row orderings simply number the cells in a matrix in a row-by-row sequence. The various paths pass through all pixels in the space but have different aims in terms of total or unit length or in linking neighbouring cells in the sequence. The most commonly used are Morton and Peano-Hilbert orderings in which the pixels are indexed according to special recursive sequences based on sub-quadrants and are shown in Figures 3.12*c,d*. They provide an addressing technique for grid units of variable size which bring about considerable savings in time when querying the database.

Compact raster structures and data analysis

Compact raster structures are efficient for data analysis when the spatial unit (pixels, lines, or polygons) are exact, static entities. Quadtrees with sufficient levels of nesting can provide as fine spatial resolution of bounded entities as vector methods of coding. Compact raster structures provide few advantages for handling the continuous fields encountered in remotely sensed images, digital elevation models, interpolated surfaces, and numerical spatial modelling because each grid cell takes a different value. They are also of little value for modelling movement or change. When data stored in block or run-length codes are used in analyses with data treated as continuous surfaces they have to be converted to the full, simple raster format. Now that computer storage space and processing speeds have greatly increased it is not so necessary today to use compact raster structures for operational coding though the methods are useful for archiving large amounts of similar data.

Summary—raster data structures. If each cell represents a potentially different value, then the simple

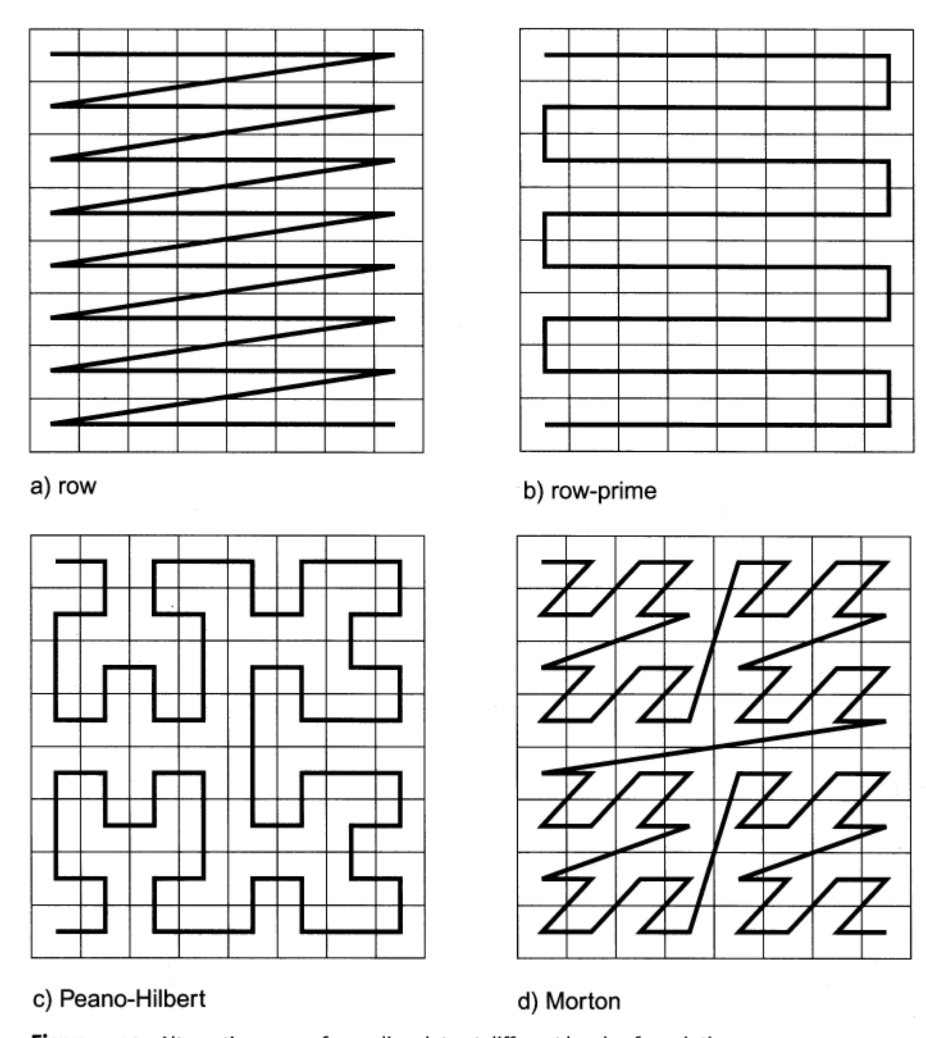


Figure 3.12. Alternative ways of encoding data at different levels of resolution

$N \times M$ array structure is difficult to improve upon. Its limitations are largely related to the volume of data and size of memory required. When 'regions' (i.e. areas of uniform value) are present, as assumed to be the case in many thematic maps, data storage requirements may be considerably reduced by using *chain codes, run-length codes, block codes*, or *quadtrees*. Run-length codes appear to be most efficient when the pixel size is large with respect to the area of the regions being displayed and sorted; as resolution improves and pixel numbers per region increase, however, block codes and quadtrees become increasingly attractive. The quadtree representation has the added advantage of variable resolution. The ease of subsequent processing varies with the data structure used.

Data organization in vector data structures

A vector database is built up from what the user perceives as a number of *feature planes* which are used to separate different classes of phenomena (shown in Figure 3.5*b*). The units are represented as crisp world objects using a coordinate space that is assumed to be continuous, not quantized as with the raster structure, allowing all positions, lengths, and dimensions to be defined precisely. In fact this is not exactly possible because of the limitations of the length of a computer word on the exact representation of a coordinate and because all display devices have a basic step size or resolution. Besides the assumption of mathematically exact coordinates, vector methods of data storage use

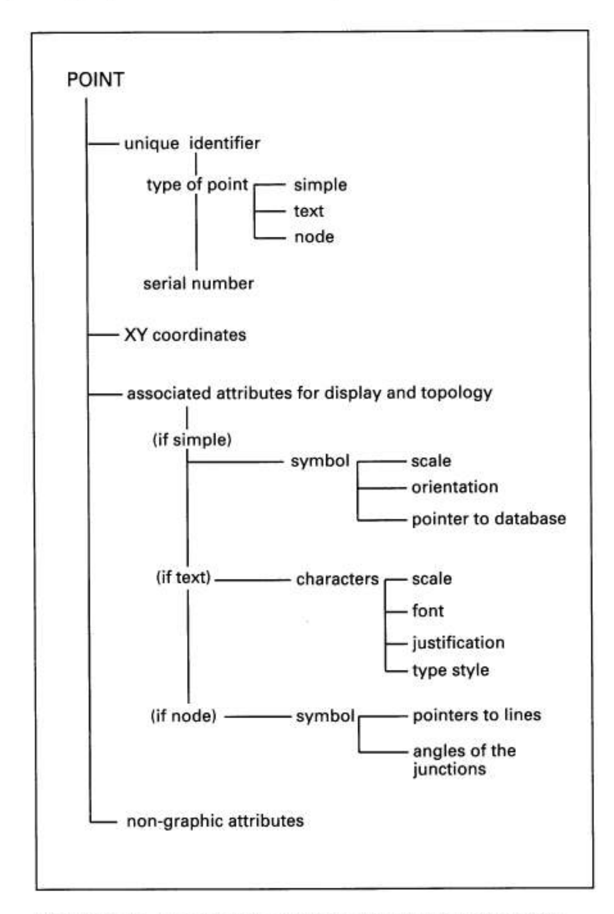


Figure 3.13. Vector data structure of a simple point entity

implicit relations that allow complex data to be stored in a minimum of space. There is no single, preferred method however. This section describes a range of vector structures used in GIS for the storage of points, lines, and areas.

Point entities Point entities may be considered to embrace all geographical and graphical entities that are positioned by a single XY coordinate pair. In addition to XY coordinates other data must be stored to indicate what kind of 'point' it is and any other information associated with it. For example, a 'point' could be a symbol unrelated to any other information. The data record would still need to include information about the symbol, and the display size and orientation of the symbol. If the 'point' were a text entity, the data record would have to include information about the text characters to be displayed, the text font (style), the justification (right, left, centre), the scale, and the orientation as well as ways of associating other non-graphic attributes with the 'point'. Figure 3.13 illustrates a possible data structure for 'point' entities.

Line entities Line entities may be defined as all linear features built up of straight-line segments made up of two or more coordinates. The simplest line requires the storage of a begin and an end point (two XY coordinate pairs) plus a possible record indicating the display symbol to be used. For example, the display symbol parameter could be used to call up solid or dashed lines on the display device even though all the segments of the dashed display were not sorted in the database.

An arc is a set of n XY coordinate pairs describing a continuous complex line. The shorter the line segments, and the larger the number of XY coordinate pairs, the closer the chain will approximate a complex curve. Data storage space may be saved at the expense of processing time by storing a number that indicates that the display driver routines should fit a mathematical interpolation function (e.g. B-splines) to the stored coordinates when the line data are sent to the display device. As with points and simple lines, arcs may be stored with data records indicating the type of display line symbol to be used.

Networks Simple lines and chains carry no inherent spatial information about connectivity such as might be required for road and transport or drainage network analyses. To achieve a linear network that can be traced by the computer from line to line it is necessary to add topological pointers to the data structure. The pointer structure is often built up with the help of nodes. Figure 3.14 illustrates the sort of data structure that would be necessary to establish connectivity between all branches of a network. Besides carrying pointers to the arcs, the nodes would probably also carry data records indicating the angle at which each chain joins the node, thereby fully defining the topology of the network. This simple linkage structure incorporates some data redundancy because coordinates at each node are recorded a total of ($n$chains + 1) times, where n is the number of chains joining a node. Attributes attached to the lines (indicated by black dots in Figure 3.14) can be used to selected preferred routes.

Polygons Polygons may be represented in various ways in a vector database. As many kinds of spatial data are linked to polygons, the way in which these entities may be represented and manipulated has received considerable attention. The following discussion is largely based on the work of Peuker and Chrisman (1975), Cook (1978), and Weber (1978), and covers several

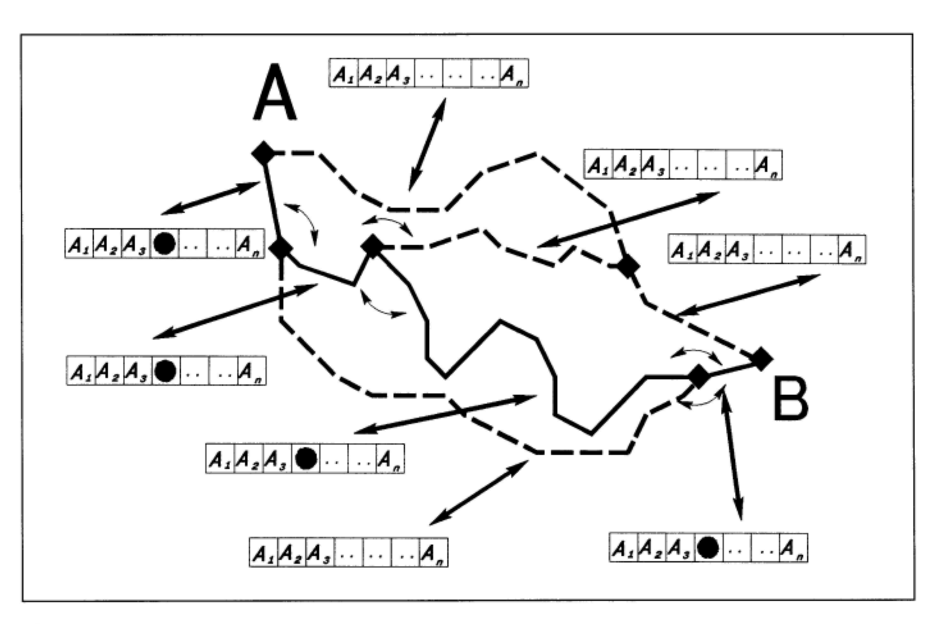


Figure 3.14. Hybrid data structure for network analysis

well-known and frequently used methods of structuring polygon data.

The aim of a polygon data structure is to describe the topological properties of areas (that is their shapes, neighbours, and hierarchy) in such a way that the associated attributes of these basic spatial building blocks may be displayed and manipulated as thematic map data. Before describing the ways in which a polygon data structure may be constructed it would be useful to state the requirements of polygon networks that geographic data impose.

First, all the component polygons on a map will each have a unique shape, perimeter, and area. There is no single standard basic unit as is the case in raster systems. Even for the most regular or regularly laid-out American street plan it will be unwise to assume that all or even some of the blocks have exactly the same shape and size. For soil and geological maps uniformity of space and size is clearly most unlikely. Second, geographical analyses require the data structure to be able to record the neighbours of each polygon in the same way that the stream network required connectivity. Third, polygons on thematic maps are not all at the same level—islands occur in lakes that are themselves on larger islands, and so on.

Simple polygons The simplest way to represent a polygon is an extension of the simple chain, i.e. to represent each polygon as a set of *XY* coordinates on the boundary (Figure 3.15). The names or symbols used to tell the user what each polygon is are then held as a set of simple text entities. While this method has the advantages of simplicity it has many disadvantages. These are (*a*) lines between adjacent polygons must be digitized and stored twice. This can lead to serious errors in mismatching giving rise to slivers and gaps along the common boundary, (*b*) there is no neighbourhood information, (*c*) islands are impossible except as purely graphical constructions, (*d*) there are no easy ways to check if the topology of the boundary

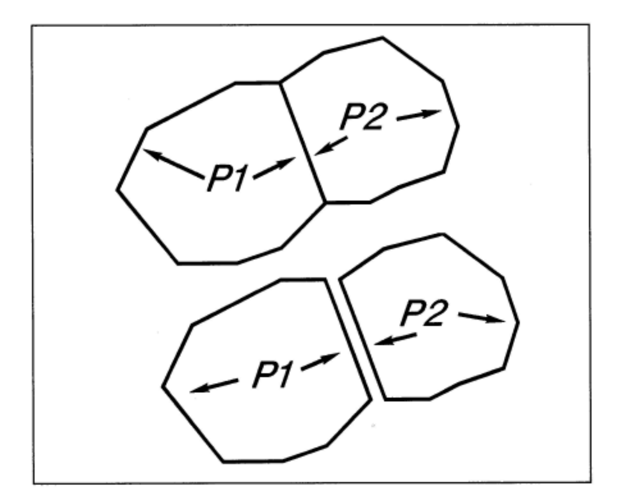


Figure 3.15. Without topology the database cannot distinguish between polygons that share boundary lines (top) or that are truly separate entities (bottom)

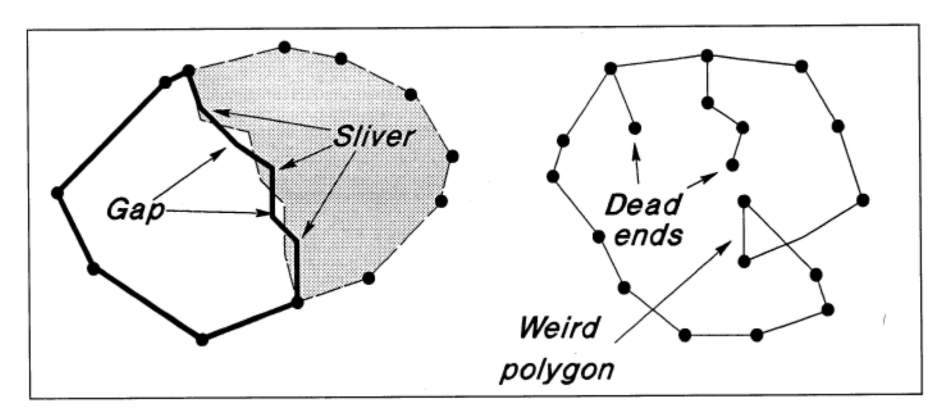


Figure 3.16. Topological errors in a polygon net

is correct or whether it is incomplete ('dead-end') or makes topologically inadmissible loops ('weird polygons')—see Figure 3.16. The simple polygon structure may be extended such that each polygon is represented by a number of chains, but this does not avoid the basic problems.

Polygons with point dictionaries With this representation all coordinate pairs are numbered sequentially and are referenced by a dictionary that records which points are associated with each polygon (Figure 3.17*a*). The point dictionary database has the advantage that boundaries between adjacent polygons are unique, but the problem of neighbourhood functions still exists. Also, the structure does not easily allow boundaries between adjacent polygons to be suppressed or dissolved if a renumbering or reclassification should result in them both being allocated to the same class. The problem of island polygons still exists, as do the problems of checking for weird polygons and dead ends. As with simple polygons, polygons may be used with chain dictionaries (Figure 3.17*b*). This has the advantage that over-defined chains, resulting from continuous or stream digitizing, may be reduced in size by weeding algorithms (see Chapter 4) without having to modify the dictionary.

Polygon attributes are linked by pointers to data tables (Figure 3.17*c*).

Polygon systems with explicit topological structures Islands and neighbours may only be properly handled by incorporating explicit topological relationships into the data structure. The topological structure may be built up in one or two ways—by creating the topological links during data input, or by using software to create the topology from a set of interlinked chains or strings. In the first case, the burden of creating the topology is thrust on the operator; the second often relies on considerable amounts of computing power; and both methods result in an increase in the amount of data that needs to be stored to describe the full polygon structure.

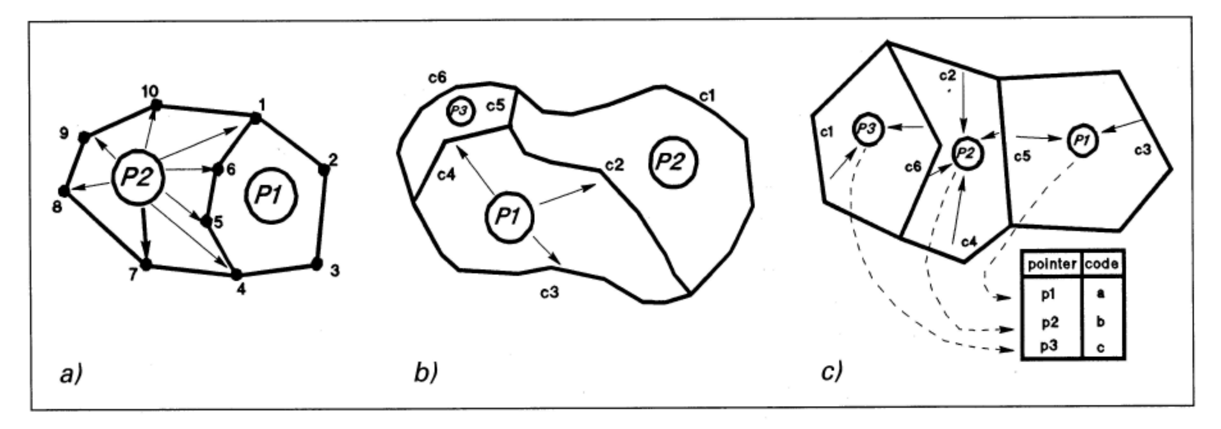


Figure 3.17. Three different ways of incorporating simple topology in polygon nets

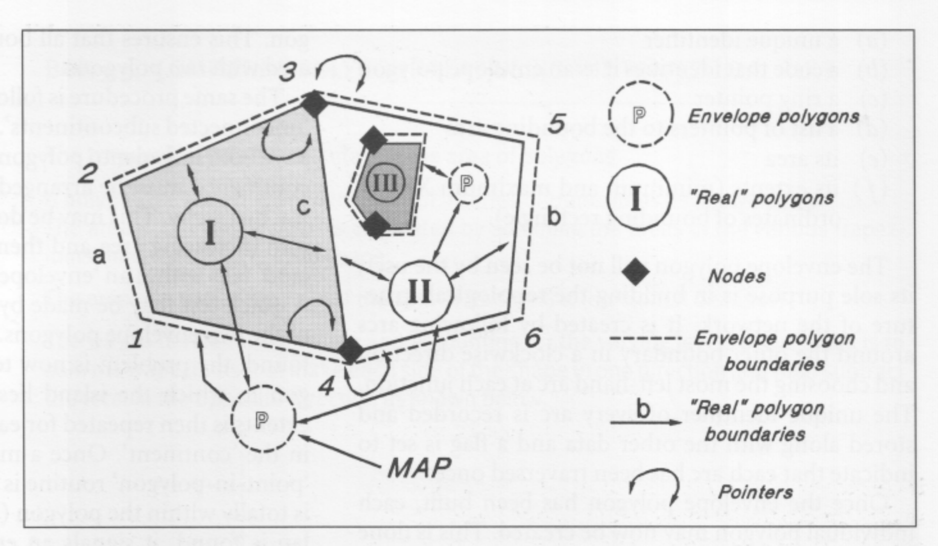


Figure 3.18. Full topological structure of a polygon map

One of the first attempts to build explicit topological relationships into a geographic data structure is the Dual Independent Map Encoding (DIME) system of the US Bureau of the Census. The basic element of the DIME data file is a simple line segment defined by two end-points; complex lines are represented by a series of segments. The segment has two pointers to the nodes and codes for the polygon on each side of the segment. As the nodes do not point back to segments, or segments point to adjacent segments laborious searches are needed to assemble the outlines of polygons. Moreover, the simple segment structure makes the handling of complex lines very cumbersome because of the large data redundancy.

A fully topological polygon network structure The structure shown in Figure 3.18 may be built up from a set of boundary chains or strings that have been digitized in any order and in any direction. The system allows islands and lakes to be nested to any level; it allows automatic checks for weird polygons and dead ends; and automated or semi-automated association of non-spatial attributes with the resulting polygons. Neighbourhood searches are fully supported. Although differing in details, the system about to be described is similar to systems used in the Harvard Polyvrt program (Peuker and Chrisman 1975).

Digitizing polygon boundaries (digitizing is described in Chapter 4) and creating polygon topology are best treated separately. The procedures used to build the boundary topology should only need to make two assumptions of the input data, namely that the polygon boundaries have been coded in the form of arcs, and that the polygon names or other records used to link the graphical to the attribute data are digitized in the form of identifiable point entities somewhere within each polygon boundary.

Stage 1. Linking arcs into a boundary network. The arcs are first sorted according to their extents (minimum and maximum X and Y coordinates occupied by the arc) so that arcs topologically close to one another are also together in the data file. This saves time when searching for adjacent arcs. The arcs are then examined to see which others they intersect. Junction points are built at the end of all arcs that join, and the arc data records are extended to contain pointers and angles. Arcs that cross at places other than the end-points are automatically cut into new arcs and the arc pointers are built.

Stage 2. Checking polygons for closure. The resulting network is then checked for closure by scanning the modified arc records to see if they all have pointers to and from at least another arc. The other arc may be the arc itself in the case of single islands defined by a single arc. All arcs failing to pass the test may be 'flagged' (i.e. brought to the attention of the operator) by causing them to be displayed in a particular way or otherwise be removed from the subset of arcs to be used for the polygon network.

Stage 3. Linking the lines into polygons. The first step in linking lines into polygons is to create a new 'envelope' polygon from the outer boundary of the map (Figure 3.18). This envelope entity consists of records containing:

(*a*) a unique identifier
(*b*) a code that identifies it as an envelope polygon
(*c*) a ring pointer
(*d*) a list of pointers to the bounding arcs
(*e*) its area
(*f*) its extents (minimum and maximum XY coordinates of bounding rectangle).

The envelope polygon will not be seen by the user; its sole purpose is in building the topological structure of the network. It is created by following arcs around the outer boundary in a clockwise direction and choosing the most left-hand arc at each junction. The unique identifier of every arc is recorded and stored along with the other data and a flag is set to indicate that each arc has been traversed once.

Once the envelope polygon has been built, each individual polygon may now be created. This is done by starting at the same place as before but this time the clockwise searches involve choosing the most right-hand arc at each junction. A tally must be kept of the number of times an arc has been traversed; once it has been traversed twice it falls out of the search. Arriving back at the starting-point, one has identified all the component lines. At the same time a check is made on the cumulative turning angle (Figure 3.18) and if this is not 360° there has been a digitizing fault and the polygon is weird. (The weird polygons should have been filtered out by the intersection-seeking step in stage 1, but if the arcs must be linked to manually entered nodes then this check is essential). As with the envelope polygon, each polygon entity receives several sets of information:

(*a*) a unique identifier
(*b*) an ordinary polygon code
(*c*) a ring pointer *from* the envelope polygon. At the same time the identifier of this polygon is written in the ring pointer of the envelope polygon
(*d*) a list of all bounding arcs. At the same time, the polygon's unique identifier is written into the record of the line
(*e*) a ring pointer *to* the adjacent polygon in the network
(*f*) minimum and maximum XY coordinates (extents) of the bounding rectangle.

The search proceeds to the next polygon in the same network at the same level in the hierarchy, and so on until all individual polygon have been built up. When the last polygon in the net has been traced its ring pointer (*e*) is set pointing back to the envelope polygon. This ensures that all bounding lines are associated with *two* polygons.

The same procedure is followed for all 'islands' and 'unconnected subcontinents'. Once all bounding arcs have been linked into polygons, the 'islands' and 'subcontinents' must be arranged in the proper topological hierarchy. This may be done by first sorting them into increasing area and then testing to see if an 'island' falls within an 'envelope' of the next largest size. A quick test may be made by comparing the extents of the two envelope polygons. Once a match has been found, the problem is now to locate the exact polygon in which the island lies. The matching of the extents is then repeated for each component polygon in the 'continent'. Once a match has been found, a 'point-in-polygon' routine is used to see if the island is totally within the polygon (see Box 3.4). If an overlap is found, it signals an error in the database or the intersection procedures (stage 1) and indicates that the operator must take remedial action. If there is no overlap, then a pointer is written from the network polygon to the envelope polygon of the enclosed island. If no overlap and no matching is found, it means that then two polygon networks are independent of each other. The ring pointer structure of envelope polygons–network polygons–island envelope polygons–island network polygons allows an infinite amount of nesting. Moreover, the nesting only needs to be worked out once; thereafter the network may be traversed easily by following the pointers.

Stage 4. Computing polygon areas. The next stage involves computing the area of the individual polygons by the trapezoidal rule (see Box 3.3). Polygons in geographical data may have many hundreds of coordinates in the bounding arcs and many island polygons (imagine the geological map of an area with lakes and drumlins) so it is usually more efficient to compute areas once, subtracting the areas of enclosed islands as necessary, and then to store this area as an associated attribute.

Stage 5. Associating non-graphic attributes to the polygons. The last stage of building the database is to link the polygons to the associated attributes that describe what they represent. This may be done in several ways. The first is to digitize a unique text entity within each polygon area, either as part of the data entry, or interactively after polygons have been formed. This text may then be used as a pointer to the associated attributes that may or may not be stored with the graphic data. The text may be used for visual display; it is linked to the polygon by the use of a point-in-polygon search (see Box 3.4). The second is to instruct the computer

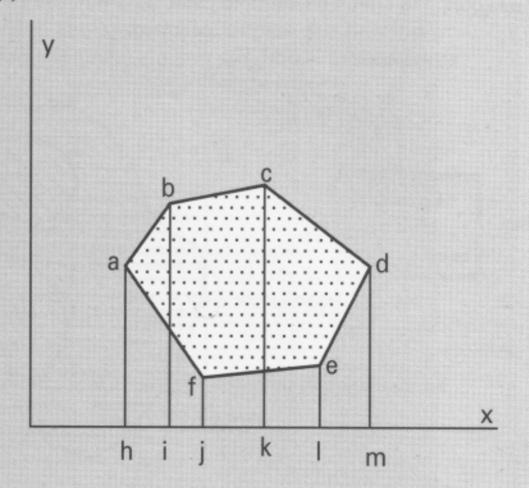
BOX 3.3. COMPUTING POLYGON AREAS USING THE TRAPEZOIDAL RULE

The trapezoidal rule for calculating the area of polygons

A polygon may be described in terms of a series of trapeziums as shown below
The area under the polygon is calculated by summing the areas of the various trapeziums that make up the total shape.

The area of a trapezium = (half the sum of its sides) × (horizontal distance)

The way to derive the total area and accounting for the varying levels of the sides is to sum the trapeziums that make each side of the shape in one direction and then subtract the total in the opposite direction as shown below.



The area of the polygon in the figure is computed by:

Add areas of upper trapeziums A, B, C, and subtract areas of lower trapeziums D, E, F. Upper trapeziums: A (h, a, b, i); B (i, b, c, k); C (k, c, d, m). Lower trapeziums: D (h, a, f, j); E (j, f, e, l); F (l, e, d, m)

to write the unique identifier of each polygon at the centre of each polygon; at the same time the computer prints a list of all polygon identifiers. This list may then be merged with a file containing the other non-graphic attributes of the polygons which may then be cross-referenced through the unique polygon identifiers.

Complex software is needed to construct the topologically sound polygon network described above but the resulting data structure of Figure 3.19 has the following advantages:

(*a*) the polygon network is fully integrated and is free from gaps and slivers and excessive amounts of redundant coordinates
(*b*) all polygons, arcs, and associated attributes are part of an interlinked unit so that all kinds of neighbourhood analyses are possible. Note that the system just described also allows the arcs to have non-graphic attributes associated with them as well
(*c*) the number of continent-island nestings is unlimited
(*d*) the locational accuracy of the database is limited only by the accuracy of the digitizer and the length of the computer word.

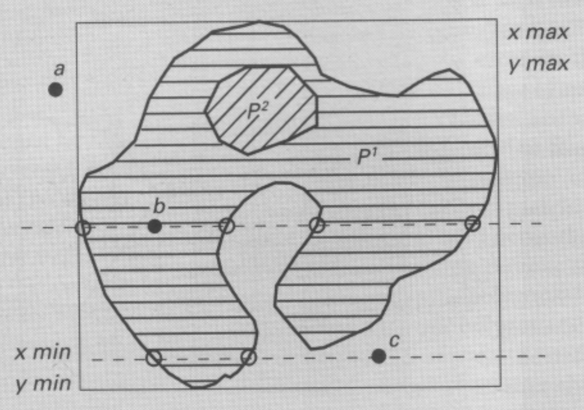
Editing and updating the polygon net Vector polygon networks may be edited by moving the coordinates of individual points and nodes, by changing the polygon attributes, and by cutting out or adding sections of lines or even whole polygons. Changing coordinates or associated attributes requires no modification to the topology and is easy. Modifying the

BOX 3.4. THE POINT IN POLYGON PROBLEM AND ITS SOLUTION

Point-in-polygon search

At least two separate steps in the creation of the polygon network involve the general problem of *point-in-polygon* search, namely the checks to see if a small polygon is contained by a larger one, and the association of a given polygon with a digitized text label. The figures below show two aspects of the point-in-polygon algorithms.

First a quick comparison of the coordinates of the point with the polygon extents quickly reveals whether a point is likely to be in or not. So point (*a*) may easily be excluded because it is outside the polygon's minimum bounding rectangle, but (*b*) and (*c*) cannot. To check if points (*b*) and (*c*) are in the polygon a horizontal line is extended from the point. If the number of intersections of this line with the polygon envelope (in either direction) is odd, the point is *inside* the polygon.



To check if an island polygon, P^2 is inside P^1, first check the extents; P^1 is then divided into a number of horizontal bands and the first and last point of each band is treated as the point (*b*) above. If the number of points for each line is odd, then the polygon P^2 is completely enclosed.

Problems may occur if any segment of a boundary is exactly horizontal and has exactly the same Y coordinate as the point X, but these may be easily filtered out. Haralick (1980) lists an alternative procedure for finding the polygon containing a point that is based on a binary search of monotone arcs—that is sequences of arcs in which the order of the vertices in the arc is the same as the order of the vertices projected onto a line *e*.

network by cutting out or adding lines and polygons requires local recalculation of topology and rebuilding the database. Consequently these kinds of data structures are not efficient for spatial patterns that are constantly changing.

SPECIAL PURPOSE VECTOR DATA STRUCTURES—THE TRIANGULAR IRREGULAR NETWORK (TIN)

An important, and much used vector polygon structure is the TIN. It is built from joining known point values into a series of triangles based on a Delaunay triangulation. The triangulation allows a variable density and distribution of points to be used which reflects the changes in attribute values within an area as shown in Figure 3.20*a*. The structure model regards the nodes of the network as primary units. The topological relations are built into the database by constructing pointers from each node to each of its neighbouring nodes. The neighbour list is sorted clockwise around each node starting at north. The world outside the area modelled by the TIN is represented by a

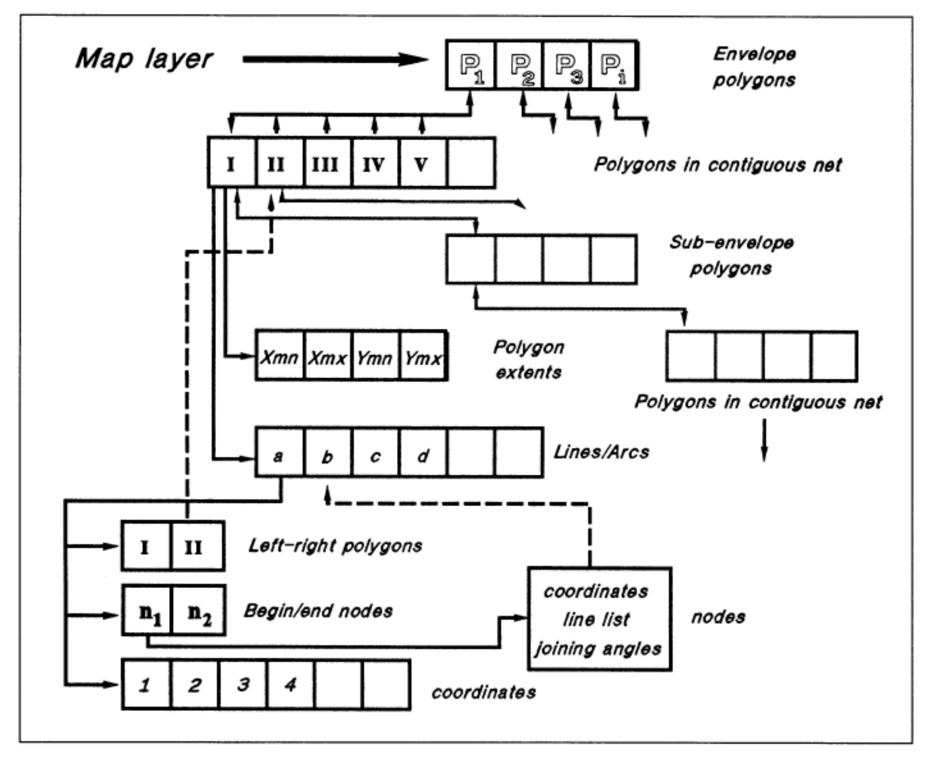


Figure 3.19. Data structure of a topologically connected polygon net

dummy node on the 'reverse side' of the topological sphere on to which the TIN is projected. This dummy node assists with describing the topology of the border points and simplifies their processing.

Figure 3.20*b* shows a part of the network data structure (three nodes and two triangles) used to define a TIN. The database consists of three sets of records called a node list, a pointer list, and a trilist (triangle list). The node list consists of records identifying each node and containing its coordinates, the number of neighbouring nodes and the start location of the identifiers of these neighbouring nodes in the pointer list. Nodes on the edge of the area have a dummy pointer set to −32 000 to indicate that they border the outside world.

The node list and pointer list contain all the essential altitude information and linkages so they are sufficient for many applications. For other applications, such as slope mapping, hill shading, or associating other attributes with the triangles, it is necessary to be able to reference the triangles directly. This is done by using the trilist to associate each directed edge with the triangle to its right. In Figure 3.20*b*, triangle T2 is associated with three directed edges held in the pointer list, namely from node 1 to 2, from node 2 to 3, and from node 3 to 1.

Nodes located in areas of greatest change help to reduce errors within the derived form. In DEMs for example values along ridges and valleys help to ensure that the resulting elevation surface does not have anomalies such as rivers which flow upstream. In using these irregularly spaced points TINs avoid the redundancies of the regular grid and provide efficient means for computing derived data such as slope.

Developments in vector data structures to improve data access times and efficient storage

The vector data model is a relatively efficient means of storing the geometric information of geographical data with only pertinent coordinate values recorded;

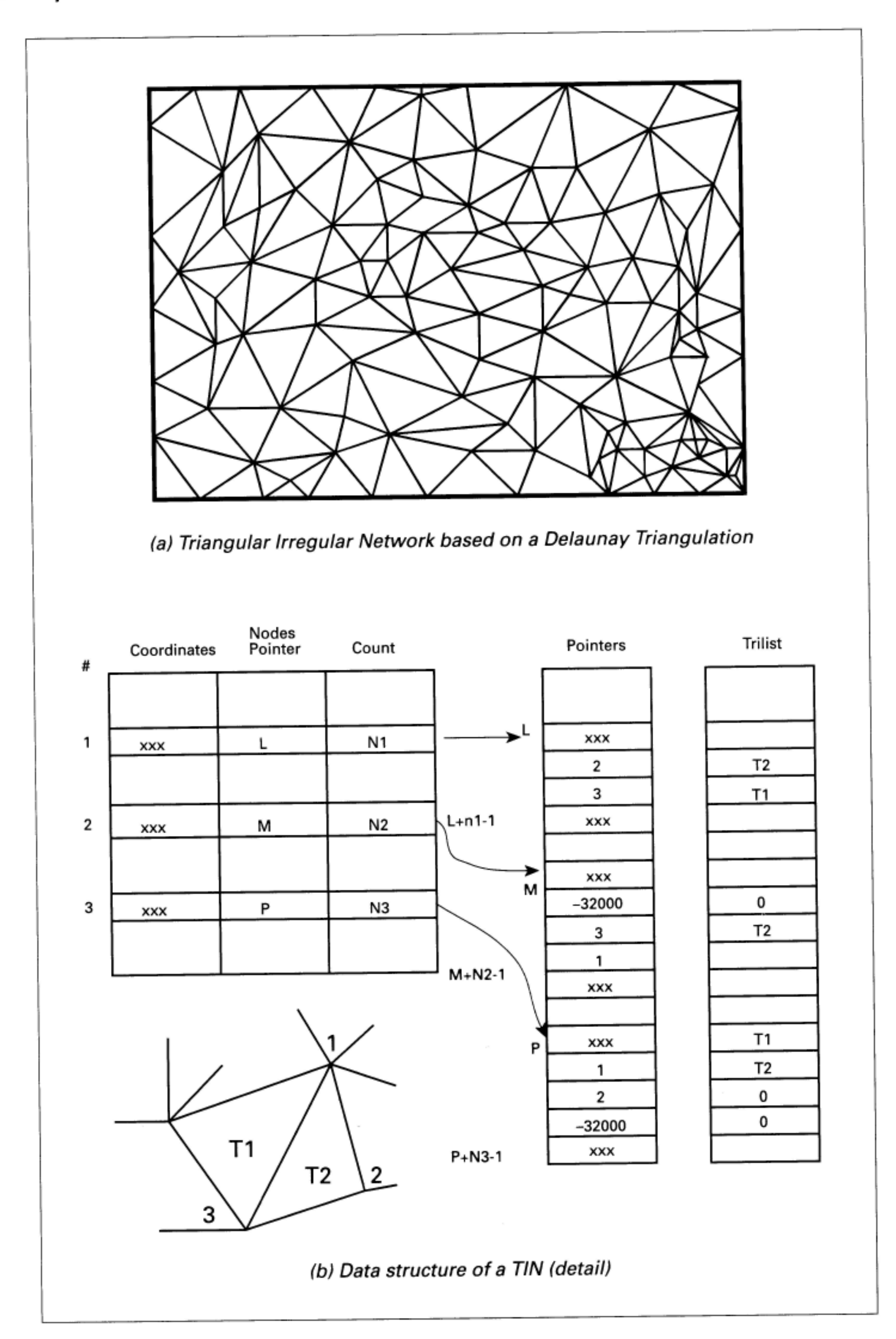


Figure 3.20. Triangular irregular networks based on the Delauney Triangulation provide a vector data model of a continuous field

the main problems have been associated with accessing the data, particularly the topological and attribute information. For example the interrelationships between the basic vector units point, line, or polygon are uniquely identified by a pointer or a label; topological structures can only be traversed by means of these unique labels. In early GIS these labels, sometimes called master index pointers, were often held in a sequential list that was the key to accessing the rest of the database. This list had two major problems; first, it was rarely fully consecutive; during editing, gaps appeared, or it increased in length, which meant that searches almost invariably were sequential. Second, the search times increased sharply with the length of the table which meant that map processing times increased non-linearly with the size of the database.

Many technical developments in GIS databases in the 1980s concentrated on the problem of making the processing of spatial data less dependent on the size of the database. The speed of computers today means that storage and accessing problems are not as noticeable as before. Changes in database organization and internal referencing have also helped to improve efficiency.

Data clustering on storage media The first attempts at improving database access times used 'brute-force' computing methods to scan the pointer arrays quickly, or to concentrate the master index array onto a small, contiguous area of disc or core storage. While this undoubtedly brought some improvements, it was at best a palliative and did little to resolve the underlying problems. Another approach involved clustering the master index pointers not only according to entity type, but also to spatial location. The data were also clustered on the disk according to geographical location.

Database indexing using B-trees and R-trees The importance of indexing a database to speed up querying was emphasized earlier in this chapter. However, the indexes for large or complex databases as in GIS applications may themselves become long and slow to query. Structures have therefore been developed which give hierarchical indexes of indexes so that searching is more efficient and directed. They are known as multi-level indexes and are particularly useful with vector data structures where much of the topological and attribute data are held in index files.

The B-tree structure provides a multi-level index which is structured using 'internal nodes' and 'leaf nodes' which are analogous respectively to the branches and leaves of a tree. In the example shown in Figure 3.21 a file of street name data is indexed alphabetically. The first index splits the data into groups of letters a–f, g–m, n–s, t–z which are listed along with pointers to the data in the respective nodes. The next level of the index splits these groups into finer divisions. When a search is initiated it is therefore limited at each stage so saving time. The B-tree structure adjusts to dynamic changes in the databases through algorithms which alter the organization following the insertion or deletion of records (discussed in more detail in Worboys 1995).

Any data type which has a linear ordering may be used as the index field in a B-tree. In GIS index files of number or text string values are useful for searches on attribute records (van Oosterom 1993). However, this structure does not address the spatial (two-dimensional) nature of many of the queries of a GIS.

There are various alternatives to the B-tree model which allow geometrical properties of the database to be included in structuring an index. For example, the R-tree (Guttman 1984) divides space into a series of boxes, known as Minimal Bounding Rectangles (MBRs). Figure 3.22 shows a series of hospital locations and the three MBRs used to divide the space. Nodes in the index represent these rectangles and a search for a particular hospital location would be directed first by an algorithm to one of these. A hierarchical structure of different rectangle sizes of MBRs may be set up using tree-shape structure for the index. The search algorithm checks which large rectangle an entity is contained in and then follows the branches of the tree down various levels until the conditions of the query are met.

Quadtrees The basic concepts of quadtrees were elucidated earlier in this chapter and their use in reducing the amount of disk space needed to represent raster data structures was described. This data structuring technique may also be used with vector based data where the presence or absence of an attribute determines the coding of a defined quadrant (Laurini and Thompson 1992). In some complex quadtree systems the cells may be referenced also in terms of the geometry of the units present such as an edge or a vertex, or number of points.

The quadtree structures used to code spatial data have been geometrically varied. In many the hierarchical partitioning of the space has been in terms of regular-shaped squares as already described. However, with some structures the data themselves have been used to determine the position and shape of the sub-

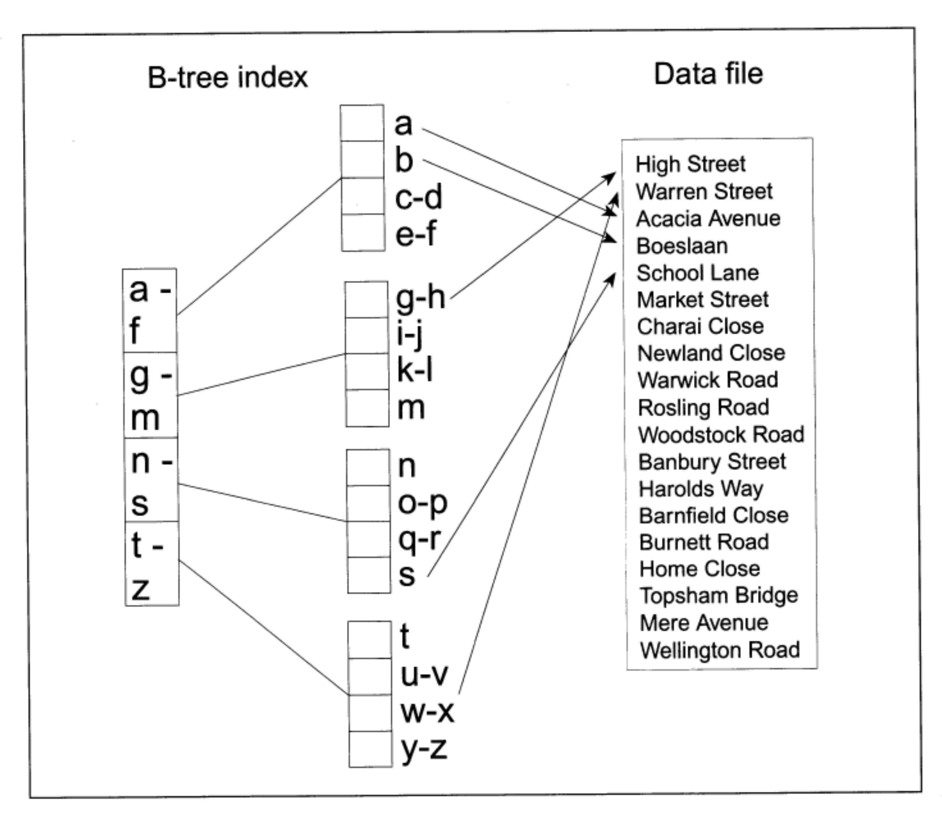


Figure 3.21. Data indexing using B-trees

divisions. The distribution or position of the points, lines, and vectors may be used to divide the space into irregular shapes at each subdivision so helping to represent the variations in density and distribution of data values at various scales (Laurini and Thompson 1992). The structural shape of irregular quadtrees is highly dependent on the time various points are inserted into it (Worboys 1995). The main problem with using quadtrees with vector data is the loss of explicit topological referencing and that the precision of point and linear features is limited.

Continuous vector coverages: tiling The real world is continuous and does not stop at the boundaries of map sheets and computer files. A seamless database simulates the continuity by linking the files for adjacent areas according to a system known as *tiling*. In theory the world is represented by an infinite set of tiles reaching in both directions. Each tile, or page as it is sometimes called, may reference a certain amount of information; extra details can be accommodated by dividing each main tile into subtiles in a manner similar to that used in quadtree structures. Tiling introduces an extra complication in that all arcs must automatically be terminated and begun at tile boundaries and topological pointers must not only reference other entities but other tiles as well. The costs are thus a larger database but one in which searching times may be kept very low by virtue of the positional hierarchy.

In principle, tiling allows limitless maps to be created and stored as only data from a few tiles need be referenced at any one time the rest being stored on disk. In practice, the sheer volume of data will exceed the financially permissible disk space, even allowing for the dramatic decrease in hard drive costs, so for country-wide mapping great reliance must be placed on networked databases or storage devices such as CD-ROMS, optical disks, or magnetic tapes for storing much of the total database until it is required.

New developments in vector data structures One of the problems encountered with conventional GIS structures is that the level of detail or resolution of the data that may be stored and displayed is fixed and determined during the input stage. When working

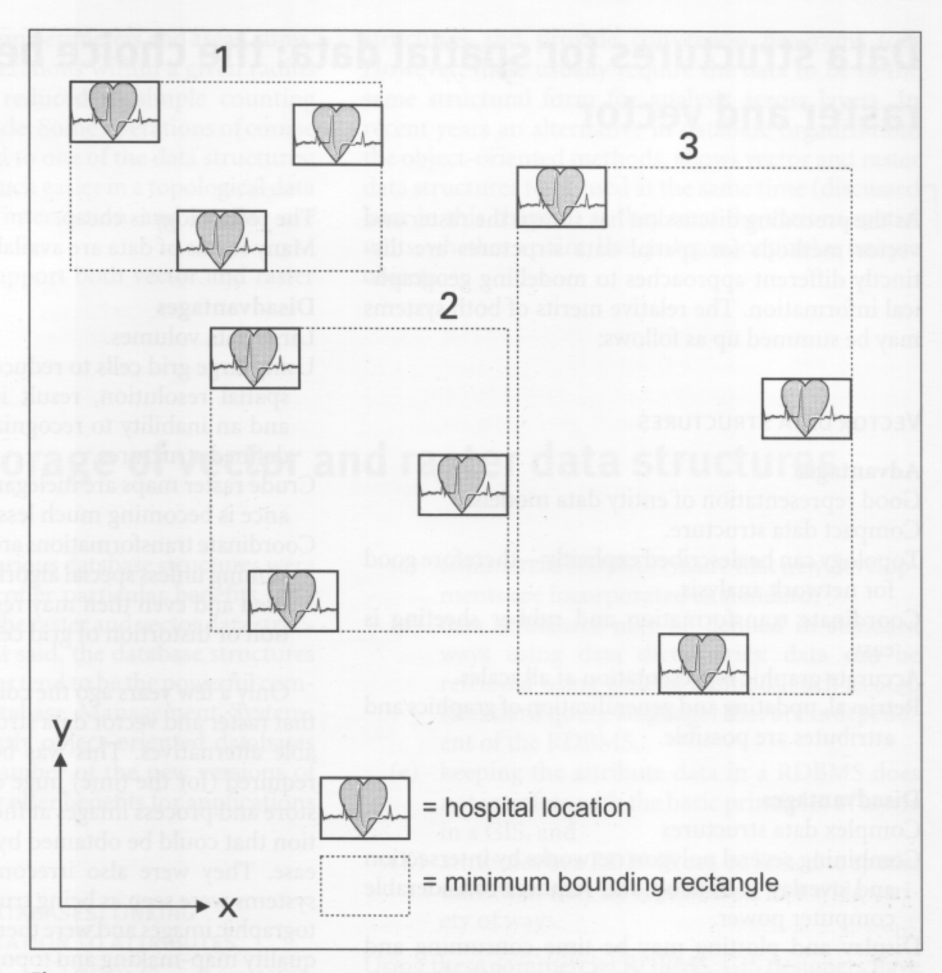


Figure 3.22. Data indexing using minimum bounding rectangles

at different map scales, the detail of the data remains the same and the user's perspective is limited by the resolution of the graphics display and the degree of zooming. Van Oosterom (1993) is one of several researchers who have developed the concept of a reactive data structure in which the level of detail and the geometric representation which the user interacts with is determined by the scale of the display. Zooming in to large-scale maps would show more detail than a small-scale representation. A linear entity might be displayed as a polygon in the former case and as a line in the latter.

Data structures for spatial data: the choice between raster and vector

As the preceding discussion has shown the raster and vector methods for spatial data structures are distinctly different approaches to modelling geographical information. The relative merits of both systems may be summed up as follows:

Vector data structures

Advantages

Good representation of entity data models.
Compact data structure.
Topology can be described explicitly—therefore good for network analysis.
Coordinate transformation and rubber sheeting is easy.
Accurate graphic representation at all scales.
Retrieval, updating and generalization of graphics and attributes are possible.

Disadvantages

Complex data structures
Combining several polygon networks by intersection and overlay is difficult and requires considerable computer power.
Display and plotting may be time consuming and expensive, particularly for high-quality drawing, colouring, and shading.
Spatial analysis within basic units such as polygons is impossible without extra data because they are considered to be internally homogeneous.
Simulation modelling of processes of spatial interaction over paths not defined by explicit topology is more difficult than with raster structures because each spatial entity has a different shape and form.

Raster data structures

Advantages

Simple data structures.
Location-specific manipulation of attribute data is easy.
Many kinds of spatial analysis and filtering may be used.
Mathematical modelling is easy because all spatial entities have a simple, regular shape.
The technology is cheap.
Many forms of data are available.

Disadvantages

Large data volumes.
Using large grid cells to reduce data volumes reduces spatial resolution, result in loss of information and an inability to recognize phenomenologically defined structures.
Crude raster maps are inelegant though graphic elegance is becoming much less of a problem today.
Coordinate transformations are difficult and time consuming unless special algorithms and hardware are used and even then may result in loss of information or distortion of grid cell shape.

Only a few years ago the conventional wisdom was that raster and vector data structures were irreconcilable alternatives. This was because raster methods required (for the time) huge computer memories to store and process images at the level of spatial resolution that could be obtained by vector structures with ease. They were also irreconcilable because vector systems were seen as being truer to conventional cartographic images and were therefore essential for high-quality map-making and topographic accuracy. Most early technical developments were undertaken in vector mode simply because this structure was most familiar to cartographers and draughtsmen. Raster systems were seen as being suitable only for overlay analysis where cartographic elegance was not required.

The previous debates on raster versus vector are no longer relevant and they have been shown to be not mutually exclusive. It has become clear that what first seemed to be an important conceptual problem is in fact largely a question of technology. The problems of graphical and hard copy presentation and data storage with raster systems have largely been overcome with high-resolution computer screens and printers, and relatively cheap media for data storage and compression techniques. In the late 1970s several workers, notabley Peuquet (1977, 1979) and Nagy and Wagle (1979) showed that many of the algorithms that have been developed for polygonal data in vector structures not only had raster alternatives, but that in some cases these were more efficient. For example,

the calculation of polygon perimeters and areas, sums, averages, and other operations within a given radius from a point are all reduced to simple counting operations in raster mode. Some operations of course still remain more suited to one of the data structures; network analyses are much easier in a topological data structure whilst map intersection and overlay are trivial in raster systems.

Today, many GIS support both vector and raster structures and provide conversion programs too. However, these usually require the data to be in the same structural form for analysis across layers. In recent years an alternative in database organization, the object-oriented methods, allows vector and raster data structures to be used at the same time (discussed later in this chapter) as they treat the various spatial units, whether a point, line, polygon, or pixel as unique objects.

Database storage of vector and raster data structures

Earlier in this chapter various database structures were described. All of them offer particular benefits (and drawbacks) for storing the raster and vector data structures just detailed. That said, the database structures used in most GIS systems tend to be the powerful commercial Relational Database Management Systems (RDBMS). More recently object-oriented databases have been used in a number of the new versions of commercial GIS as they offer benefits for applications in a number of fields.

Hybrid relational databases: linking geometric representation to attributes

The availability of commercial RDBMS, such as INFO, ORACLE, INGRES, INFORMIX, and similar products, greatly eased the work of GIS system designers as they were able to apply ready developed and tested systems to their data handling needs. These databases allowed designers to divide the problems of spatial data management into two parts. The first part was how to represent the geometry and topology of the spatial objects—should this be done using vector or raster data structures? The second part was how to handle the attributes of the spatial objects, which may be done using the commercial RDBMS. The resulting hybrid structures (sometimes referred to as georelational models) have a number of distinct advantages:

(*a*) attribute data need not be stored with the spatial database but may be kept anywhere on the system, or even on-line via a network,

(*b*) attribute data can be expanded, accessed, deleted, updated without having to modify the spatial database,

(*c*) commercial RDBMS ensure that new developments are incorporated as standard,

(*d*) data structures may be defined in standard ways using data dictionaries: data can be retrieved using general methods such as SQL (standard query language) that are independent of the RDBMS,

(*e*) keeping the attribute data in a RDBMS does not interfere with the basic principles of layers in a GIS, and

(*f*) attributes in a RDBMS can be linked to spatial units that may be represented in a wide variety of ways.

Using these commercial RDBMS, GIS designers have created a variety of hybrid structures including the following:

1. ARC-NODE—RDBMS. This is probably the most used database system in which the full vector arc-node topology is used to describe networks of lines and polygon boundaries as explained above. Each spatial unit is identified by a unique number or code and it may be placed in a chosen layer or overlay. The attributes of the spatial unit are stored as records in relational tables that may be handled by the RDBMS. Some of the topological information and attributes such as coordinates, neighbours, areas, coordinates of minimum bounding rectangles, and similar data may also be stored in tables in the RDBMS. Figure 3.23*a* shows a typical example of such a data structure.

One complication with such a system is the creation of new spatial entities and their associated attributes when layers are intersected because this means building new sets of records and links in the RDBMS.

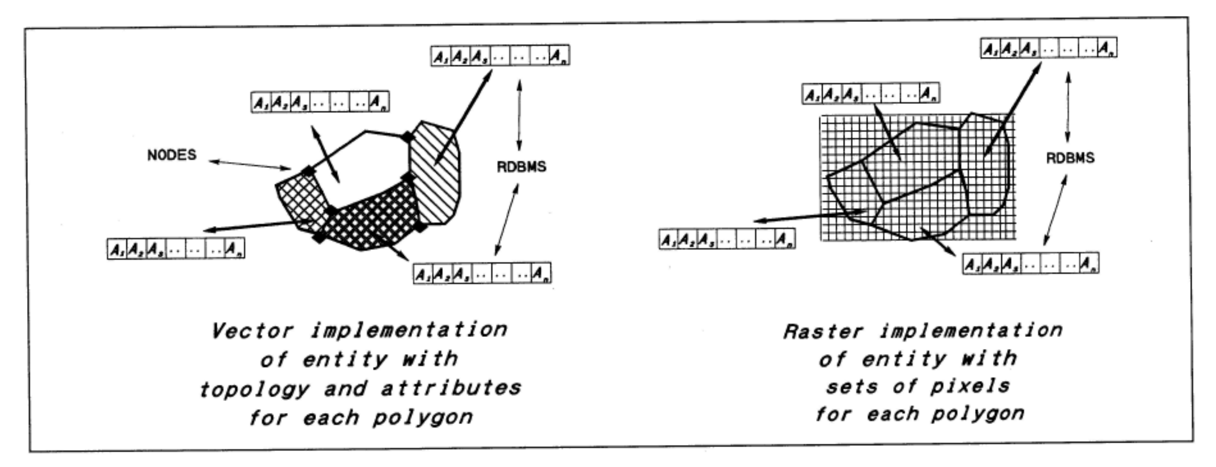


Figure 3.23. Equivalent hybrid vector (a—left) and raster (b—right) data structures for modelling crisp polygons

2. Compact raster—RDBMS. If the spatial objects are represented by *sets of pixels* instead of topologically linked lines then the raster equivalent to the above may be created. When most of the spatial data refer to thematic units that are internally homogeneous, such as choropleth map units, then raster compaction methods of run-length codes may be used to save space. Figure 3.23*b* shows an example of the simple raster hybrid approach for homogeneous polygons.

3. Quadtree—RDBMS. Quadtrees may also be used for data compression but because they allow the data to be represented at various levels of spatial aggregation they may permit different levels of spatial resolution to be used in different layers. This could be useful when intersecting data on narrow river valleys with broad land planning units, which could be done in vector mode, but would be difficult in raster systems that permit only a single cell size or level of resolution.

4. Object—RDBMS. In recent years, an object-oriented approach (discussed earlier) has been adopted in organizing both raster and vector data structures in the same GIS. In these systems the various geometric and attribute data are stored in relational tables (Gahegan and Roberts 1988) and object-oriented programming languages provide analytical functionality as well as a graphical object-based interface to the data. The systems allow the benefits of object-oriented organization of geographical data to be exploited within the well-known relational database environment.

Object-oriented database structures: unifying attribute and geometric storage

Object-oriented databases require geographical data to be defined as a series of atomic units. This obviously favours data defined using the entity conceptual model. Geographical data are characterized by a series of attribute and behavioural values which define their spatial, graphical, temporal, and textual/numeric dimensions (Worboys 1994). Part of the attribute definition will describe the geometric nature (point, line, polygon, or cell) of the object; more than one geometric type may be used to reflect the differences in shape found at different spatial scales.

The attribute and behaviour variables are themselves object classes for which their properties and the methods used on them are defined. Hierachical relationships may be set up with the various classes; for example the 'arc' object may be a subclass of the 'polygon' object. Topological links between various object instances and classes are established explicitly through

object pointers and operators such as 'direction', 'intersection', 'adjacent-to', 'overlaps', 'left-of', or 'right-of'.

Spatial data organization in object-oriented databases has proved attractive to certain GIS users as it offers a way of modelling the semantics and processes of the real world in a more integrated, intuitive manner than possible in relational systems (Kidner and Jones 1994). People working with human-made objects, such as utility companies, have found that these systems provide an approach which suits the data types they use and the querying capabilities needed. Hierarchical structuring and the representation of relatively complex relationships between object classes may be controlled directly so giving flexibility in database updating and changing.

The structuring of the database into a series of self-contained, fundamental units brings with it both problems and possibilities. It is very difficult to break down continuous spatial fields into separate units. How do you break up a hill into a series of distinct objects that may be used in analysis and modelling? The choice of boundary is often subjective (see Chapter 11 and Burrough and Frank 1996). However where data do support this and an objective discretization may be made, the possibilities of data exchange are increased. The objects may be used in new applications even where a different structuring is required. While reusable object libraries involve a considerable investment of time and money, they have already been shown to give a considerable return (Worboys 1995). Spatial data object libraries may now be accessed across the Internet.

To date the implementation of object-oriented databases in GIS has been limited. The problem is that there are few generic object-oriented database products available to act as an engine to support GIS functionality.

The debate on relational-hybrid GIS versus object orientation

Recent developments in object-oriented databases has encouraged debate on their merits with respect to existing relational-hybrid systems. The main points may be summed up as follows:

Relational-hybrid GIS

Advantages

Modifiability of spatial data after inputting to the system.
Data retrieval and modelling functionality is provided by the DBMS.
Easy data integration from other systems particularly for the attribute data.
All aspects of the data are stored in a specialized file structures.
Ease of use.
Sound theoretical foundation for relational database.

Disadvantages

Poor handling of temporal data.
Coordinate data tend not to be subject to the rigorous database management as might be applied to attribute data, so issues of security and integrity exist.
Relies on the spatial position or attribute value for querying or modelling.
Slow handling of querying especially when dealing with complex objects.
Querying and modelling limited to functionality provided by the GIS (or data must be exported).

Object-oriented database GIS

Advantages

The semantic gap between the real-world objects and concepts and their representation in the database is less than with relational databases.
The storage of both the state and the methods ensures database maintenance is minimized.
Raster and vector data structures may be integrated in the same database.
The data exchange of objects is supported.
Fast querying of the database especially when complex objects and relationships have to be dealt with as fewer join operations are needed.
Requires less disk space than relational entities which need to store many more index files.
Enables user-defined functions to be used.

Disadvantages

There is no universally accepted object-oriented model so different database products have different standards and tend to be tied to one particular O-O language.

Identifying objects is often difficult, particularly in continuous spatial surfaces.

Requires the definition of functions and topology as well as objects.

Limited application of indexing because of the incompatibility of it with the notion of encapsulation and object-identity.

No established standards such as SQL and provisions for a general query language or query optimization is made difficult by the complexity of the object types in the system.

There is less theoretical and practical experience with O-O approach than the hybrid method.

(After Arctur and Woodsford 1996, Graham 1994, Herring 1992, Milne *et al.* 1993, Worboys 1994, Worboys 1995.)

Recent GIS developments have been aimed at giving users more flexibility in defining spatial units and providing more analytical techniques. This should allow them to concentrate more on the spatial relationships between the variables rather than on their computer representation and the limits imposed by database structures. New developments in interoperability and Open GIS (discussed in Chapter 12) will help.

Questions

1. Discuss the limitations that computer coding imposes on spatial data structuring.

2. Why are database management systems so important? What are their main functions?

3. Think of the main digital geographical data that you come across in day-to-day living. What data models are used to represent the information and why?

4. Why is it important for the user to be aware of the database structure when using a GIS?

5. Review the different methods used for speeding up data access and compression. Consider a range of different GIS applications where these techniques are important.

6. Design an object-oriented and a hybrid relational database for one or more of the following applications:

   - an environmental study of an oil pipeline spill
   - an archaeological investigation of a prehistoric settlement
   - a police department
   - a utility company (water, gas, electricity, telephones)
   - a holiday resort
   - a soil survey organization.

   What data would you use and how would you define them in the database? Think about the limitations and benefits of each approach.

Suggestions for further reading

Holroyd, F., and Bell, S. B. M. (1992). Raster GIS: Models of raster encoding. *Computers and Geosciences*, 18: 419–26.

Jacobsen, I. *et al.* (1992). *Object-oriented Software Engineering*. Addison-Wesley, Wokingham.

Laurini, R., and Thompson, D. (1992). *Fundamentals of Spatial Information Systems*. Academic Press, London.

Worboys, M. F. (1995). *GIS—a Computing Perspective*. Taylor & Francis, London.

FOUR

Data Input, Verification, Storage, and Output

Building an accurate GIS database of spatial entities is an exacting task. Raw geographical data are available in many different analogue or digital forms, such as maps, aerial photographs, satellite images, or tables. There are three, not mutually exclusive ways to create a digital geographical database: (*a*) acquire data in digital form from a data supplier, (*b*) digitize existing analogue data, and (*c*) carry out one's own digital survey. In all cases the data must be geometrically registered to a generally accepted and properly defined coordinate system. Whether in analogue or digital form, the data need to be converted to the internal database structure of the GIS being used. With existing digital data sets this often involves using a standard exchange format between the supplier and client systems. Where the original data are in analogue form the coordinates of the entities are recorded digitally using devices such as digitizers, scanners, and stereoplotters. Once data have been captured they must be checked for mislocation and value errors. Most GIS provide data editing tools for this work. Attribute values must also be linked to the entity database, which involves building links to relational tables or spreadsheets. Essential details concerning the provenance and other important characteristics of the data are stored in metadata files. Large geographical databases are not necessarily stored on a single computer but may be distributed over a network which is accessible to many users. It is important to store a GIS database safely because much effort goes into creating it. Magnetic and optical media provide data storage for back-up and distribution of data products. All or part of a spatial database can be displayed ephemerally on computer screens or on hardcopy paper or film.

Sources of geographical data

Creating a GIS database is a complex operation which may involving data capture, verification, and structuring processes. Because raw geographical data are available in many different analogue or digital forms, such as maps, aerial photographs, satellite images, or tables, a spatial database can be built in several, not mutually exclusive ways. These are:

- acquire data in digital form from a data supplier,
- digitize existing analogue data, and
- carry out one's own survey of geographic entities
- interpolate from point observations to continuous surfaces.

In all cases the data must be geometrically registered to a generally accepted and properly defined coordinate system and coded so that they can be stored in the internal database structure of the GIS being used. The desired result should be a current, complete database which can support subsequent data analysis and modelling.

For many people the most common source of geographical data is the paper or digital topographic or thematic map, which is a graphical representation of the distribution of spatial phenomena. Maps are drawn to a certain scale and show the attributes of entities by different symbols or colouring. The location of entities on the earth's surface is specified by means of an agreed coordinate system.

Images derived from optical and digital remote sensing systems mounted in aircraft and satellites provide much spatial information over many levels of temporal and spatial resolution (see Box 4.1). Stereo aerial photographs are overlapping, analogue images having many applications including the creation of topographical maps and orthophotomaps by *photogrammetry* (see Chapter 5). Stereo aerial photographs are a major source of data for the human interpretation and mapping of geology, soil, vegetation, or land cover, and they are also valuable background documents for placing other spatial data in a proper geographical context. Digital photogrammetry and digital orthophoto mapping provide data on terrain elevation and land cover directly in digital form without the need for conversion from a paper analogue document (e.g. Plate 1.1).

A wide range of scanners mounted in satellites or aircraft provide digital data directly. There are systems that are passive receivers from reflected radiation, and the amount of reflected energy is recorded for an increasingly large range of wavelengths, including thermal and microwave regions of the electromagnetic spectrum. Passive systems react to the different levels of absorption and reflectivity of the components of the earth's surface, thereby providing information on spatial patterns of different kinds. Active systems include side-band radar scanners, laser altimeter scanners (e.g. Plates 3.5, 3.6) and sonar scanners (mounted in hovercraft or boats for under water applications) that provide information on surface elevation and surface material densities. The use of the data they deliver is determined by the type of sensor and the temporal and spatial resolution, which depends on the path and altitude of the supporting platform.

The traditional means of collecting geographical data is by ground or field surveys to record sample values at known locations, using instruments ranging from questionnaires and soil augers to automated chemical probes (Stienstra and van Deen 1994). The results of these surveys are usually recorded in terms of a series of point location and attribute values in a table; these may be interpolated to a continuous surface using methods given in Chapters 5 and 6. Today, the recording of values is made easier through digital data loggers, essentially mini-computers attached to measuring instruments or hand units, which record values at set time intervals or at the command of the user. These may then be read either directly or through an exchange format into a computer database such as a GIS.

Georeferencing

It is most important that all spatial data in a GIS are located with respect to a common frame of reference. For most GIS, apart from local studies, the common frame of reference is provided by one of a limited number of geodetic coordinate systems.

The most usual and convenient coordinate system used in GIS is that of *plane, orthogonal cartesian coordinates*, oriented conventionally north–south and east–west. Because the earth is not flat, but nearly spherical, the *longitude* or east–west position is related to the Greenwich meridian and the *latitude* or north–south position is related to the Equator. Since the earth is not even a true sphere but flattened at the poles, geodesists have devised several *ellipsoids* for mapping the true curved surface of the earth on to a plane of which the most common are the International, the

BOX 4.1. THE DETECTION OF REFLECTED RADIATION BY REMOTE SENSING

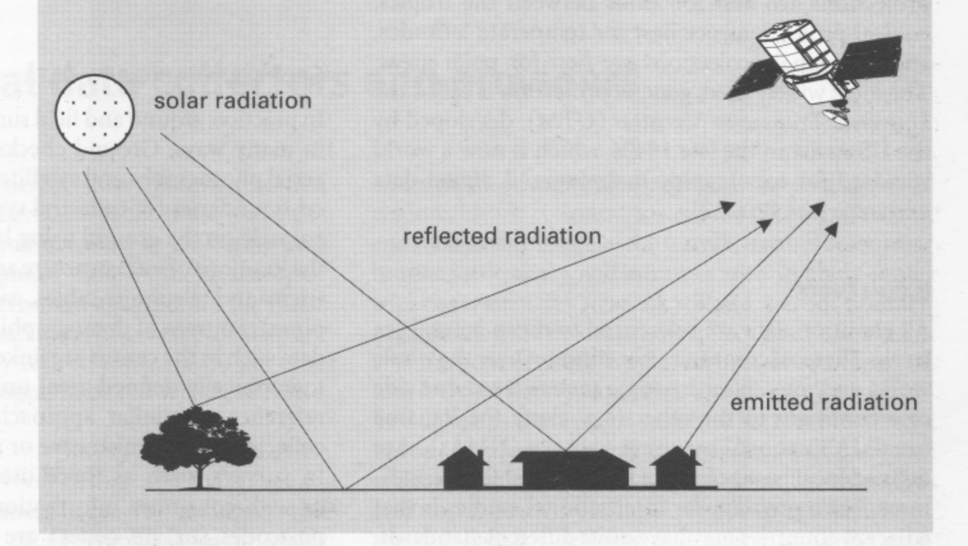
Remote Sensing Systems

Remote sensing is the collection of data about an object without coming into contact with it. This involves the detection and recording of values of emitted or reflected electromagnetic radiation (energy emitted by all bodies with a temperature greater than −273°C) using sensors onboard aircraft and satellites.

The data record the amount of radiation reflected or emitted by features on the earth's surface and may be either in analogue (as with aerial photography) or digital form. With the latter, data are recorded as a series of grid cells (*pixels* or picture elements) each coded with a value representing the radiation detected by the sensor from the area of the earth's surface covered by the pixel. Simple systems only register a single value for each of a limited number of wavebands (a range of wavelengths of electromagnetic radiation) that have been chosen to give as much information as possible about certain aspects of the earth's surface such as vegetation, rock and soil minerals, and water. For example, the scanners on the French SPOT satellite record values for four wavebands (Band 1; 0.5–0.6 μm, Band 2; 0.6–0.7 μm, Band 3; 0.7–0.8 μm, Band 0.8–1.1 μm) in order to be able to detect differences in water, vegetation, and rock. Multispectral scanners now in development record continuous spectra for each pixel and therefore generate huge amounts of data.

The spatial resolution, or the area covered by a single pixel, depends on the altitude of the sensor, the focal length of the lens or focusing system, the wavelength of the radiation, and other inherent characteristics of the sensor itself. Pixel sizes vary from a square kilometre for data from meteorological satellites to a few square centimetres for aircraft-based, high-resolution sensors.

Data collected by remote sensing are affected by atmospheric conditions and irregularities of the platform, such as tilt and orientation. Geometric and radiometric corrections are needed prior to data input to the GIS to minimize these distortions. The visual appearance of the images can be improved by increasing contrast, stretching the range of grey levels or colours used, and by edge detection, to make it easier to recognize spatial features.



BOX 4.2. COMMUNICATIONS BETWEEN COMPUTERS OVER THE INTERNET.2

Details of the UTM

The UTM uses the following ellipsoids: International Spheroid, Clarke 1866 (Africa), Clarke 1880 (North America), Everest, or Bessel (Maling 1992).

The projection is the Gauss–Kruger version of the Transverse Mercator.

The projection is only intended for mapping between 84 °N and 80 °S.

The unit of measurement is the metre.

The UTM divides the world east–west into 60 zones of longitude each 6° wide. Zones are numbered from west to east with zone 1 having its western edge on the 180° meridian.

The UTM divides the world north–south into 20 zones of latitude, starting at the equator. Zones are 8° high, except the most northerly and southerly, which are 12° high.

Each zone has its own coordinate system.

The Eastings of the origin of each zone is given a value of 500 000 m. For the southern hemisphere the equator is assigned a value of 10 000 000; for the northern hemisphere the equatorial value is 0.

Krasovsky, the Bessel, and the Clarke 1880 ellipsoids (Brandenburger and Gosh 1985).

There are three main ways for projecting locations from an ellipsoid onto a plane surface, namely cylindrical projections, azimuthal projections, and conical projections (Maling 1992). The best projection to use depends on the location of the site on the surface of the earth, and cartographers find that cylindrical projections are best for lands between the tropics, conical projections are best for temperate latitudes, and azimuthal projections are best for polar areas. The most widely used, general projection is called the *Universal Transverse Mercator* (*UTM*), developed by the US Army in the late 1940s, which is now a world standard for topographic mapping and digital data exchange (see Box 4.2).

Base levels

All elevation data are referenced to mean annual sea levels. These, of course, are not constant over the whole world and may differ by some metres from one side of a continent to the other (e.g. along the Panama canal). All national mapping agencies (NMAs) have defined local base reference levels to suit their conditions, but a problem for international studies is that adjacent countries may have quite different standards.

Georeferencing and GIS

It is essential for GIS and spatial analysis that all data are referenced to the same coordinate system. When working within a single country one usually adopts the coordinate conventions without question, but when building multinational data sets it is very important to ensure cross-border equivalence of ellipsoid, projection, and base level.

Georeferencing raw data

In practice, ground and field surveys are georeferenced in many ways. Ground checks are needed to locate aerial photographs and satellite images correctly. Parcel boundaries for cadastral systems are surveyed accurately on the ground using laser theodolites, as are the locations of infrastructure and utilities (roads, electricity and telephone cables, gas, water, and sewerage pipes). Sources of demographic and socio-economic data such as the census are linked to less accurate, cartographically defined areal units as the basic spatial reference. A similar approach is adopted for data collected by administrative or municipal authorities. In surveys, such as those used in market research or with consumer information such as store cards, postcodes (or zip codes) are used to indicate the

location of a data record. In surveys of the natural world geographical position is usually recorded in terms of latitude/longitude or with respect to a national grid.

Georeferencing with GPS

The problem of defining and recording the location of a data point has been eased through the development of Global Position Systems (GPS) which are being used increasingly in many types of data collection exercise, often in tandem with data loggers. These instruments are able to define the geographical location and altitude, to varying degrees of accuracy, anywhere on the earth's surface using triangulation geometry based on signals emitted by the NAVSTAR GPS satellites. A hand-held ground receiver must be able to receive signals from at least three of these satellites (Figure 4.1) that provide details of their orbits and an atomic clock correction. The distances from each satellite to the receiver are a function of the number of whole wavelengths and the phase shift (Fix and Burt 1995) and this information, combined with the positional information, is sufficient to compute the location of the receiver in latitude/longitude or grid reference, and altitude. These results are displayed on the GPS handset and may be downloaded into a computer system. GPS is an important source of locational information particularly in areas where map coverage is limited. The main limitations are the accuracy

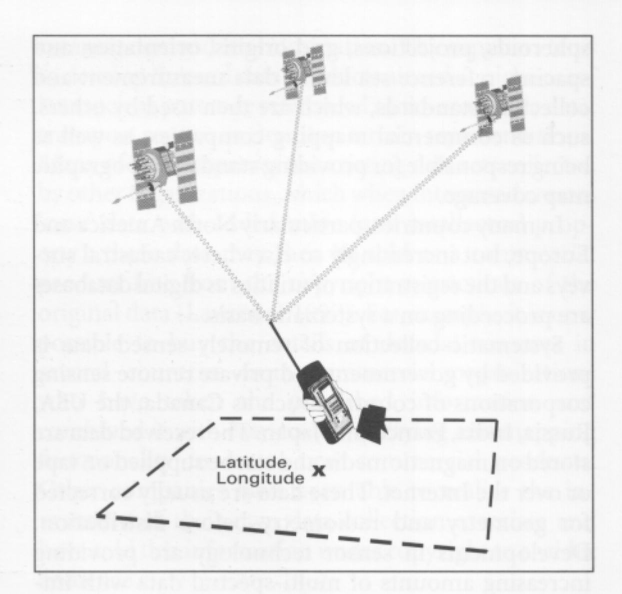


Figure 4.1. Global Positioning Systems are used to obtain latitude, longitude, and altitude data anywhere in the world

with which the geographical position can be derived because the precise time code is dithered by the US Department of Defense, the number of satellites in view, and the quality of the GPS receiver. By using a local base station on a well-located object such as a block of flats or a lighthouse, *differential GPS* measurements can improve spatial resolutions to within 1 m accuracy (Kennedy 1996).

Geographical data collectors and providers

National Mapping Agencies (NMA), natural resource survey institutes, commercial organizations, and individual researchers are all involved in collecting and disseminating geographical data, both in analogue and digital form. Government agencies are the main collectors, providers, and users of geographical information; of all data they collect, some 60–80 per cent may be classified as geographical (Lawrence 1997).

Systematic data collectors

Traditionally, government agencies maintain the systematic collection of geographical data. Many countries have national or regional mapping agencies responsible for collecting systematic data on phenomena such as the nature of the terrain, natural resources, human settlements, and infrastructure. In many countries these agencies were, or still are, part of military organizations. The data they collect tend to be general purpose and are employed by a broad range of users. Other surveys are carried out by more specialized agencies that record variables such as land ownership, employment and journey to work patterns, soils and geology, rainfall and temperature, river flow and water quality. National mapping agencies have the mandate for defining and maintaining the map

spheroids, projections, grid origins, orientation and spacing, reference sea levels, data measurement and collection standards, which are then used by others, such as commercial mapping companies, as well as being responsible for providing standard topographic map coverage.

In many countries, particularly North America and Europe, but increasingly so elsewhere, cadastral surveys and the registration of utilities as digital databases are proceeding on a systematic basis.

Systematic collection of remotely sensed data is provided by government and private remote sensing corporations of countries such as Canada, the USA, Russia, India, France, and Japan. The received data are stored on magnetic media and can be supplied on tape or over the Internet. These data are usually corrected for geometry and radiometry before distribution. Developments in sensor technology are providing increasing amounts of multi-spectral data with improved spectral, radiometric, spatial and temporal resolution, giving more detailed information. The recent launch of ERS2, and the forthcoming Canadian Radarsat as well as forthcoming earth-observation satellites from France, Japan, and the USA will supplement these data sources. The main problems associated with remotely sensed data as a systematic data source stem from (*a*) their availability which is limited by factors such as cloud cover, (*b*) their cost (variable between the different government and commercial supplying organizations), and (*c*) the need for special image processing systems so that recognizable entity-based information can be derived from the raster images.

Ad hoc data collectors

Data collected by commercial organizations such as private surveyors, civil engineers, market researchers, political organizations, or academic institutions are often project specific. They are collected for a specific purpose such as a market research exercise, a mining evaluation, an environmental impact assessment, or a university project. Although the kinds of entities and attributes recorded may cover myriad subjects, these surveys use the basic coordinate systems and base mapping information provided by systematic survey. The scale of the study and the observation methods used, along with the data classification and interpretation, are often quite unique to the survey, thereby limiting their use in other applications. Access to the data is often restricted because of commercial interests.

Data providers

Data providers offer or sell geographical data in a variety of detail, formats, scales, and structures. Traditionally data collectors and providers were the same organizations but recently these roles have become more separate and distinct as some organizations, including commercial companies, have acquired the rights to resell data they have themselves not collected. For example, there are agencies in many countries which are distributors of remotely sensed data from the US Landsat and French SPOT systems.

Geographical information available for distribution by the data providers is subject to various physical and legal restrictions, which are especially stringent where military agencies have been responsible for collecting the information. Policies on data ownership vary greatly between countries, and licensing and copyright regulations restrict the dissemination of both government and commercially collected data. Further restrictions to access stem from institutional policies of pricing to recover some of the costs of collection and distribution (cf. Burrough and Masser 1997).

Until recently, most geographical data were supplied in analogue form and this was often a major hurdle for GIS users because the effort they had to make to convert paper maps to digital files was immense and often as great as resurveying an area using contemporary instruments. Over the last decade, however, government and commercial organizations have realized the market potential for digital geographical data, and in some countries at least, such data can be easily obtained, albeit for a price and with copyright restrictions. It is now possible to obtain or purchase a broad range of digital geographical data for both natural and socio-economic applications. The costs of acquiring these data vary greatly between countries, with states such as the USA not charging for data collected using public funds while NMAs in other lands are required to charge market prices.

Much digital data provided by governments is of a general nature—elevation, roads, towns, rivers—for a broad range of users, so distribution efforts are supported by a large market. There have also been projects in the United States, Europe, and Australia to develop digital databases which integrate map data for a number of different scales or themes. These offer a useful data product which is able to meet various user demands. For example Lytle *et al.* (1996) describe the development of a national digital soil database (STATSGO) containing maps of soil organic carbon,

available water capacity, soil depth, and other properties. Kineman (1993) describes a five-year interagency project sponsored by the US NOAA and EPA to develop a global ecosystems (spatial) database to help in the investigation of global change that include data on vegetation cover boundaries, surface climatology, soils, elevation, and terrain. Langas (1997) describes efforts to create a multinational GIS for the Baltic catchment. These databases are aimed at providing information that may be used in a range of applications. In the last few years there have been discussions in many countries concerning the development of integrated national, continental, and global geographical databases. These are in varying stages of implementation and are discussed in more detail in Chapter 12.

In parallel with these developments in the public sector, commercial organizations have started to develop digital geographical datasets to help meet the increasing demand for information. The suppliers tend to develop themed data products aimed at niche markets such as postcode referenced socio-economic data, transport routes, or elevation terrain representations. Many of the companies use existing data sets collected by other organizations, which when integrated creates 'new' data with added value (and intellectual property); these data may then be sold as a commercial product free from the copyright restrictions of the original data (Lawrence 1997). For example, it is now possible to buy the details of decadal censuses in several countries from commercial organizations who have taken the published information and combined it with easy-to-use querying and mapping functionality to create new commercial products. Other companies have been able to realize the investment of their own data collection exercises and have sold this information under copyright to a wider audience.

Acquiring digital datasets from a data supplier

Digital data sets from data suppliers range from small-scale, country and continent-wide coverages available for nothing on the Internet to expensive, tailored products geared to specialist market sectors (Burrough and Masser 1997). For many local government and business users they provide an essential source of digital data that they can rely on and mean that these users do not have to cope with the overheads of digitizing or collecting their own data.

Although using existing datasets is attractive, serious attention must be paid to data compatibility when data from different suppliers are combined in one project. There may be differences in projection, scale, base level, and description of attributes that could cause problems. For example, in a survey of desert terrain a nomad is likely to classify the area differently from a geological engineer. Water-bearing capabilities, shade, and grazing productivity are likely to dominate the interpretation of the landscape features by the nomad, whilst economic potential or geological hazards might be the concern of the engineer. Users therefore need to ensure that when using an existing data set, its semantics correspond with theirs. This is not always easy to ascertain, as many data sets do not include any supporting information, known as *meta data*, about the way the survey was conducted and the information collected.

At a more practical level a user will need to consider the following characteristics of the data to ensure that they are compatible with the application:

- the currency of the data,
- the length of record,
- the scale of the data,
- the georeferencing system used,
- the data collection technique and sampling strategy used,
- the quality of the data collected,
- the data classification and interpolation methods used,
- the size and shape of the individual mapping units.

Where data are used from a number of sources, and particularly where the area of study crosses administrative boundaries, difficulties in data integration are caused by different geographical referencing systems, data classification, and sampling strategies of the individual surveys. Users need to be aware of these

problems which are particularly prone when compiling inter-state and international data sets (e.g. Mounsey 1991, Lytle *et al.* 1996, Burrough and Masser 1997).

Once the compatibility issues have been addressed the next stage involves the transfer of the data from the source to the GIS. Digital data distribution is relatively easy using media such as DAT tapes, CD ROMS, and floppy disks and the Internet is becoming increasingly important. Where necessary the data must be converted from the encoding and structuring system of the source to that of the GIS to be used. Converting data from one geographical data handling system to another has always been difficult though essentially a technical issue with problems stemming from the widely varying computer transfer formats the many different suppliers of digital geographical data use. Most GIS vendors use a variety of commercial database management systems to handle the attribute data efficiently whilst using their own solutions to handle the geometric and topological aspects of the data (as discussed in Chapter 3). Only the simplest of interfaces for data exchange used to be provided. These proprietary solutions meant when transferring data from one system to another all spatial entities encoded on supplying system had to be degraded to basic sets of points and lines ('spaghetti data') which could then be read into the new GIS where the topology and data structures had to be rebuilt (as shown in Figure 4.2*a*).

In recent years these problems have begun to be addressed through sensible agreements on technical standards for transferring the data. Today there are various nationally and internationally defined standards for data exchange coming from both governments and the GIS vendors. Commercial transfer formats such as DXF and E00 have come to be generally supported by the data providers and GIS vendors for the importing and exporting of data. The data may be exported from the local database structure of one system to this format and then imported to another system, which is then able to read the files and convert it to its own system, as shown in Figure 4.2*b*. For example many GPS allow data to be exported directly into a GIS using one of these formats. Standardization in GIS data is being driven by governments and international bodies such as the US NSDI (National Spatial Data Infrastructure—National Research Council 1994) and CEN (Commission Européan Normalisation—David *et al.* 1996). Several countries have established transfer formats such as the NTF (National Transfer Format) in the UK which provide

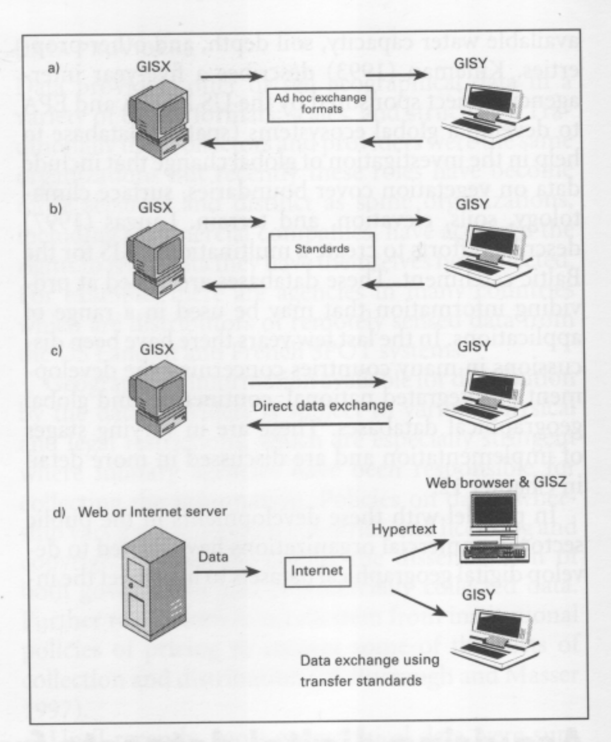


Figure 4.2. Transferring digital data between different systems

specifications for exchanging the data as well as for recording the basic quality of the component information (Guptill 1991).

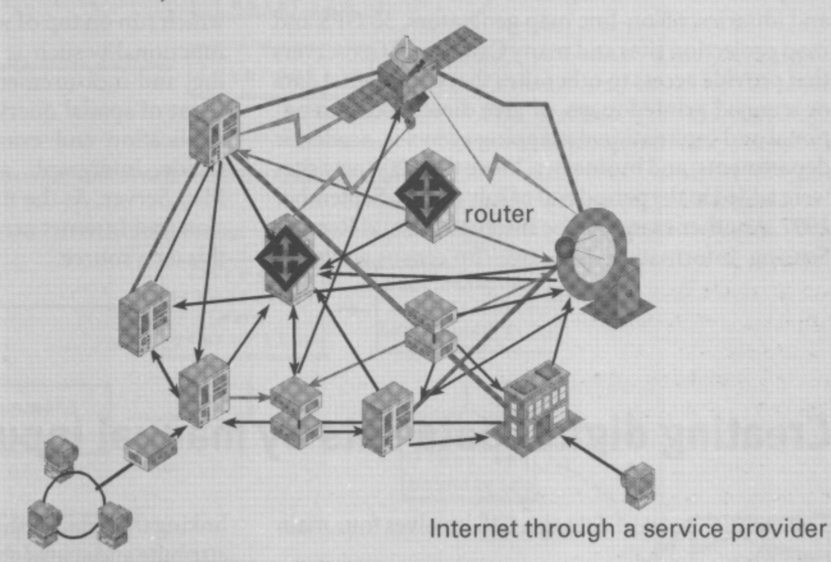
In further improvements, some commercial GIS will accept data directly from other systems without the need for exchange files as shown in Figure 4.2*c*. This is likely to improve further thanks to discussions and research in a number of parts of the world on the concept of 'Open GIS' which would allow data to be transferred easily between systems (discussed further in Chapter 12). This will ease the bottleneck of data inputting and encourage greater use of existing geographical data.

With digital remotely sensed data or scanned aerial photographs, the information is already held in pixel form, but this may be not compatible with the format of either a raster or a vector GIS. Various kinds of preprocessing are needed, such as adjusting the resolution, the pixel shape (both for orientation and distortion), and the cartographic projection in order to ensure topological compatibility with the database (detailed in Chapter 5). Other preprocessing activities include classifying groups of pixels as land cover types such as urban, corn fields, olive groves, etc. Raster-vector conversion algorithms may also be used

BOX 4.3. THE INTERNET AND INTRANET

The Internet and Intranet

The *Internet*, a public enterprise, is a vast digital network of computers. They are joined by an array of different data transmission media such as satellite and radio links, and fibre optic, unshielded-twisted pair, co-axial, and telephone communication lines. Connecting these media involves an array of different devices ranging from complicated routers and data switches, through simple signal amplifying hubs to modems. These all transfer the data using a standard coupling protocol known as TCP/IP (Transmission Control Protocol/Internet Protocol). A user accesses the Internet either directly from their computer which is joined to one of these network links, or from dialling into (using a form of telephone communication line) a commercial Internet service provider. This is shown schematically below:



By accessing computers at other locations the Internet performs a number of roles:

- as a communication mechanism through electronic mail and discussion groups.
- as an information access tool
- as a file transfer tool (using FTP—File Transfer Protocol)
- as a terminal to run programmes on computers in another location (using TELNET)

The *World Wide Web* (WWW) is a term used sometimes synonymously with the Internet. It is in fact the main information tool of the Internet through which data, written using hypertext media onto 'web pages', are accessed. These are available to the connected world through a 'web server' and viewed by users at remote locations using 'web browser' software. For the GIS user the WWW provides data, and is a source of information about the different technologies, about current academic research including on-line journals, and even about employment opportunities

The *Intranet* is a local, private linked network which provides all the facilities of the Internet but limits the usage to those directly connected to it. It is used predominantly within organizations for disseminating information and for communication, and because it is separate from the public network data and connection security are ensured.

(described in more detail later in this chapter) to generate linear and polygonal entities. The resulting mapped images may then be transferred using a exchange format to a general GIS.

New methods of capturing data in a GIS have been brought with the Internet and Intranet (see Box 4.3). Whole libraries of vector, raster, and object data are being offered now on the Internet as well as directory information on different datasets. To take one example, as of September 1997, Roelof Oddens Bookmarks (see http://kartoserver.frw.ruu.nl/html/staff/oddens/oddens.htm) listed more than 2,350 sources of new digital maps, more than 50 sources for the Netherlands, 20 interactive routing programs (see Chapter 7) for road navigation, 124 Electronic Atlases and libraries, 65 on-line map generators, 30 GPS and map projection sites and many Carto- and Geoservers that provide access to other sites that offer digital data or scanned printed maps, or give direct access to national and international mapping agencies, academic departments, and businesses. More than 200 new sites were added in the period end of July to mid-September 1997. Another example, the Internet GIS and Remote Sensing Information site (ftp://ftp.census.gov/pub/geo/gissites.txt), provides a comprehensive list of online GIS and Remote Sensing sites, arranged in alphabetical order. Many of the data are free (for example the 1 km resolution Digital Chart of the World) though increasingly costs are being imposed by the data providers through licensing agreements required to access the datasets. Not all data are up to date, some have historical value, and others provide daily or weekly information.

Web browsing software is used to access data stored at particular provider sites (web servers) and either used 'on-line' using proprietary GIS software as shown in Figure 4.2*d* or downloaded to the user's local computer (known as the client) using a standard transfer format. The proprietary software consists of programs which run on top of web browsers and offer basic GIS functionality such as data integration, spatial browsing, and measurements. Some support the development of spatial querying capabilities within another application and examples include Autodesk MapGuide, Intergraph, and ESRI MapObjects Internet Map Server. At the moment speed of data transmission and Internet access are the main limits to using this data source.

Creating digital data sets by manual input

The manual input of data to a GIS involves four main stages:

- entering the spatial data,
- entering the attribute data,
- spatial and attribute data verification and editing,
- and, where necessary, linking the spatial to the attribute data.

Figures 4.3 and 4.4 summarize the processes for raster and vector data structures. The various database structures used in GIS also require the data to be input differently. The main differences are associated with the second and third stages of the process. With the hybrid relational arc databases the spatial and attribute data are stored separately within the GIS (see Chapter 3) and need to be linked prior to any analysis. With the object-type relational databases a similar procedure is needed as the spatial and attribute data are stored in different databases and the object identifier is used in linking the data. With the object-oriented database the attribute and spatial data are linked through object and class definitions (see Chapter 3). There is no need for a separate linking process to take place.

Entering the spatial data

With the entity model, geographical data are in the form of points, lines, or areas/pixels which are defined using a series of coordinates. These are obtained by referring to the geographical referencing system of the map or aerial photograph, or by overlaying a graticule or grid onto it.

The simplest way of inputting data is then to type the coordinates into a file or input program of a GIS. The enormous labour of writing down coordinates and then typing them into a computer file may be greatly reduced by using hardware devices such as digitizers, scanners, or stereoplotters to encode the X and Y coordinates of the desired points.

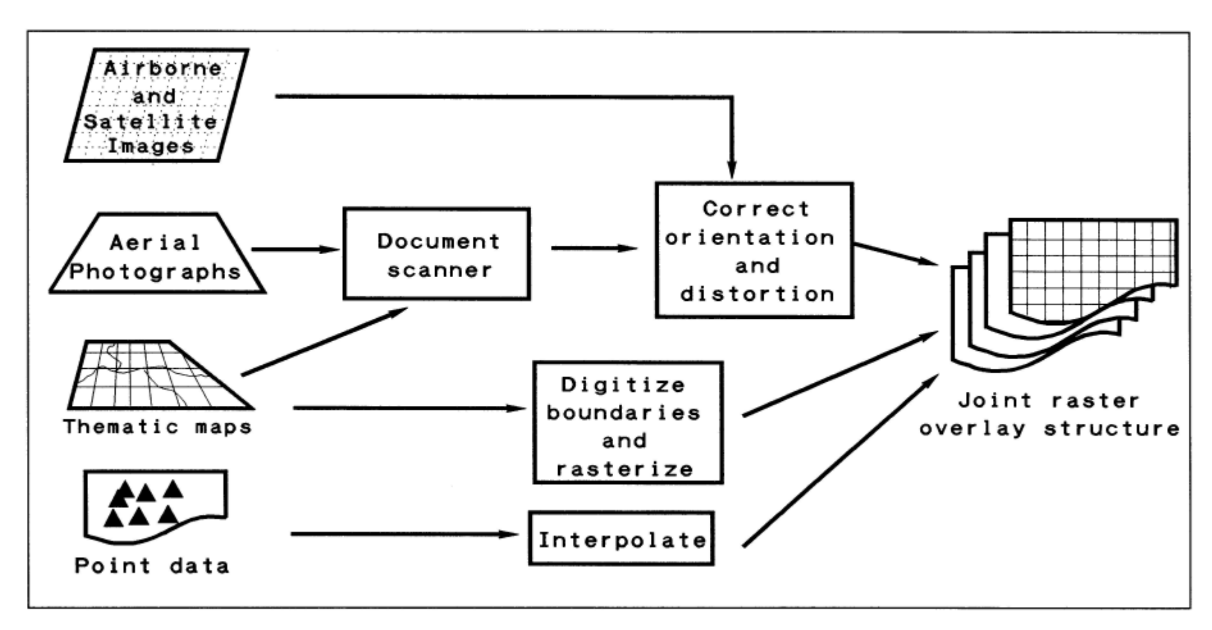


Figure 4.3. The capture and processing of spatial data to build a raster database

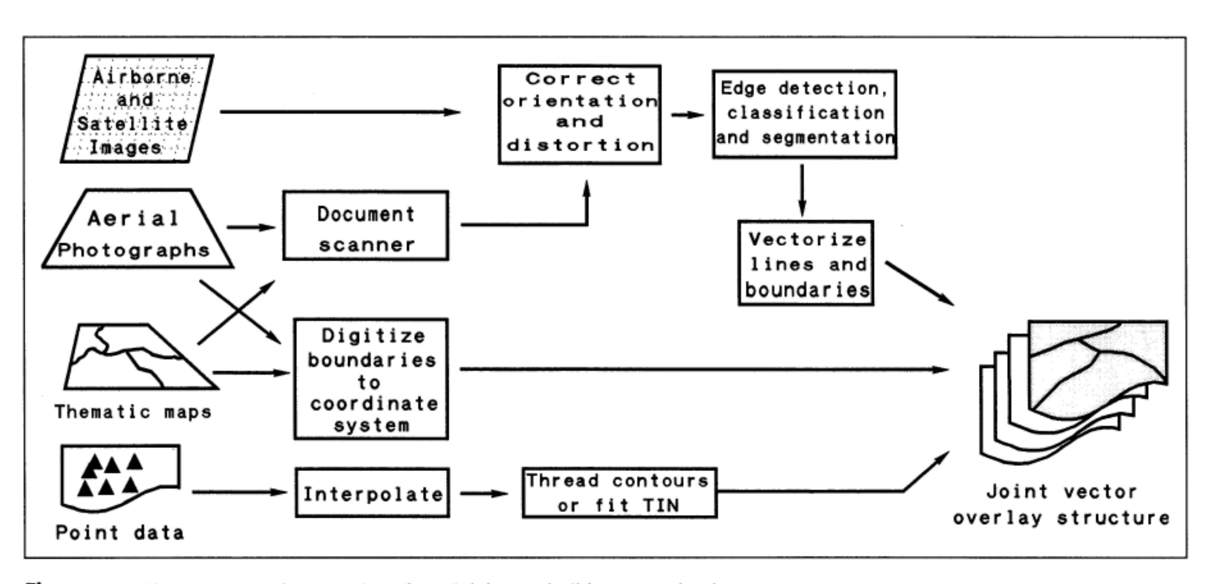


Figure 4.4. The capture and processing of spatial data to build a vector database

Digitizers A digitizer is an electronic or electromagnetic tablet upon which a map or document is placed (shown in Figure 4.5). Embedded in the table, or located directly under it, is a sensing device that can accurately locate the centre of a pointing device which is used to trace the data points of the map. The most common types currently used are either the electrical-orthogonal fine wire grid or the electromagnetic and range in size 30 × 30 cm (12 × 12 inches) to approximately 1.1 × 1.5 m (40 × 60 inches).

The pointer may be held in a cursor-device such as a mouse, or a puck or alternatively in a pen-like appliance; these may be either corded or cordless. Positioning the cursor or pen over a point on the map

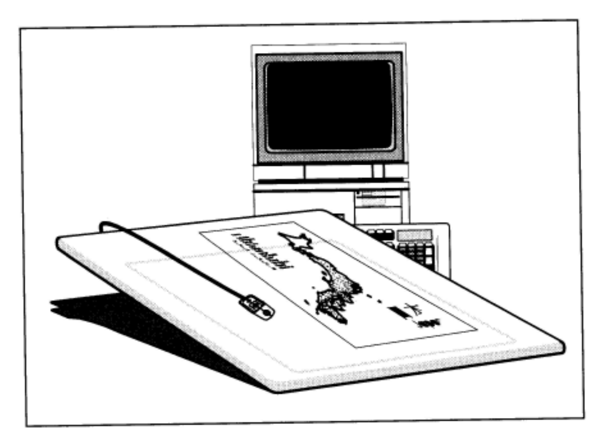


Figure 4.5. A digitizer tablet for manually inputting spatial data

and pressing a button on it, sends an electrical signal directly to the computer indicating the cursor's coordinates with respect to the digitizer's frame of reference. Where considerable accuracy is required, a puck consisting of a coil embedded in plastic with an accurately located window with cross-hairs is used. Pucks usually have 4, 12, or 16 buttons for program control, so that the operator can add additional information or coding to the database, such as identifying labels to the points, lines, or areas, whilst digitizing.

The principal aim of the digitizer is to input quickly and accurately the coordinates of points and bounding lines. The map to be captured is secured to the digitizer surface with tape and the exercise begins by digitizing at least four known points which bound the map region. These act to fix a framework within which all subsequent coordinates are recorded, and may then later be adjusted for alignment and scale. These reference points are converted to the absolute map coordinates by simple scaling routines either at this initial stage or when the data capture exercise is completed.

Lines may be digitized in two ways, known respectively as stream and point digitizing. In stream digitizing, the cursor is placed at the beginning of the line, a command is sent to the computer to start recording coordinates at either equal time intervals or equal intervals in the X or Y direction and the operator moves the cursor along the line taking care to follow as closely as possible all the bends and undulations. At the end of a line or at a junction, the computer is instructed to stop accepting coordinates. The rate at which coordinates are recorded is dependent on the speed with which the operator can trace the line and usually relatively fewer coordinates will be recorded for straight line sections than for bends and other intricate parts.

The main disadvantage with stream digitizing is that if the operator does not work at the rate expected, too many coordinates may be recorded. Greater locational errors are also to be expected because the operator has less time to position the cursor accurately. For this reason, particularly when the operator has sufficient skill to be able to choose the best places to digitize, many people prefer point digitizing. In this mode the operator instructs the computer to record every coordinate by pushing a button on the puck.

Digitizer accuracy is limited by the resolution of the digitizer itself, by the skill of the operator and, it should also be remembered, by the accuracy of the original data. A digitizer having a specified resolution of 0.0254 mm (0.01 inch) will not be able to add any accuracy in the capture process to a map of hand-drawn boundaries that have been traced with a 0.4 mm pen. Resolutions offered by some of the most sophisticated instruments are quoted as 400 lines per mm (10 160 lines per inch) with an accuracy of ±0.05 mm (Calcomp WWW page). Digitizing, in spite of modern table digitizers, is time-consuming and enervating work, a drudge. Digitizing fatigue will affect performance and it is not wise to spend more than four hours per day behind a digitizer if consistent work is to be maintained.

Many GIS have digitizing programs as part of the software. Where they are not provided, various commercial digitizing programs are available which allow the data to be exported in a standard transfer format for inputting to the GIS.

Rasterization Rasterization (vector to raster conversion) is the process of converting vector data into a grid of pixel values. This involves basically placing a grid over the map and then coding the pixels according to the occurrence or not of the phenomena. Rasterization capabilities are provided by a number of GIS. Pavlidis (1982) provides a comprehensive discussion of these algorithms and of their advantages and disadvantages.

Document scanners Scanners are devices for converting analogue data into digital grid-based images. They are used in geographical data capture to convert paper maps to high-resolution raster images which may be used directly or further processed to give vector representations. They work on the principle that markings on a map reflect a light beam differently from areas that are blank; these differences in intensities are recorded digitally (using up to 32 bits per cell for high-quality digital orthophotos) to give a digital

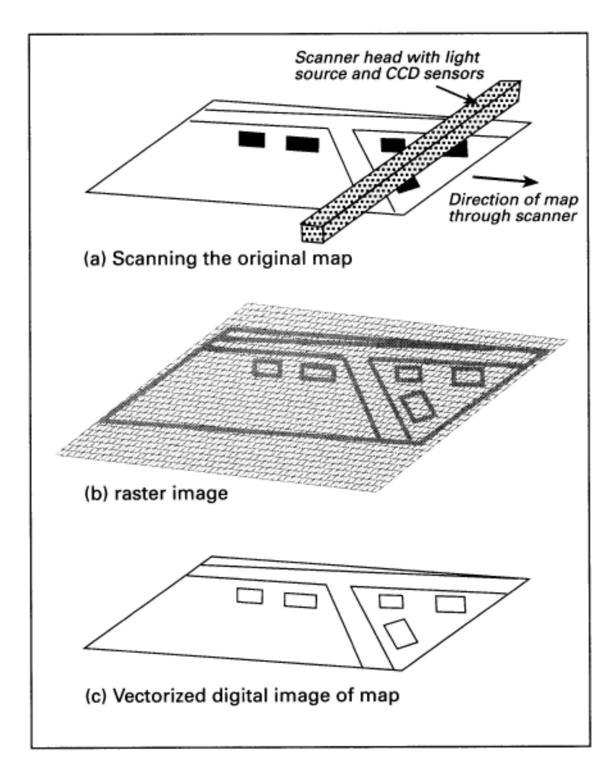


Figure 4.6. The actions of a raster document scanner. (*a*) Scanning the original map; (*b*) Raster image; (*c*) Vectorized digital image of map

image made up of a grid of reflection values. Such images are very similar to the raster images obtained by remote sensing.

There are two basic kinds of scanners, namely those that record data on a step-for-step basis, and those that scan a whole document in one operation in a manner akin to xerography. The first kind of scanners incorporate a source of illumination on a movable arm (usually light emitting diodes or a stabilized fluorescent lamp) and a digital camera with a high-resolution lens. The camera is usually equipped with special sensors known as Charge Coupled Devices or CCDs arranged in an array. These are semiconductor devices that translate the photons of light falling on their surface into counts of electrons which are then recorded as a digital value (Almelio 1974).

A digital two-dimensional image of the map is built up by the movement of either the scanner or the map (as shown in Figure 4.6). The map to be scanned can be mounted either on a flat bed, or on a rotating drum. With flatbed scanners, the light source is moved systematically up and down over the surface of the document. For larger maps up to 1.5 m (60 inches) wide, scanners are used which are mounted on a stand and the illumination source and camera array are fixed in position. The map is then moved past them by a feeding mechanism, similar to a fax machine. With drum scanners, the map is mounted onto the outside of a rotating drum and the sensor is moved across the map at right angles to the direction of rotation.

With these photo-mechanical scanners the cell or pixel size of the scanned image is controlled by the step size and the rate at which either the document or the light source moves. The scanning resolution may be set by the user to suit the application, with high-quality scanners supporting up to 1800 dots per inch (dpi). There are also scanners such as the Laserscan VTRACK system in which the laser beam can be trained to follow clearly distinguishable features (see next section).

Modern document scanners that use a method akin to xerography resemble laser printers in reverse because the scanning surface is manufactured with a given resolution of light-sensitive spots that can be directly addressed by software. There are no moving parts, except a movable light source and the resolution is determined by the geometry of the sensor surface and the amount of memory rather than by a mechanical arm. Colour scanning is possible by using a range of light-sensitive materials for each pixel. Both the mechanical and the xerographic methods yield large data files: to give the reader a little idea of the volumes of data generated, the scanned bit file of only the contours on a 30 × 50 cm topographic map sheet of a mountainous land scanned at 1600 dpi includes some 11.5 megabytes of data.

The scanned image will be far from perfect even with the best possible scanners, for it will contain all the smudges and defects of the original map plus mistakes caused in areas where the map detail is complex. The digital image may then need interactive beauty treatment to make it usable with operations to remove excess data from the image using thresholding or binarization (classifies all pixels from a fixed range of grey values to one level to create a binary image) and filtering to remove isolated pixels. If the original map quality is poor giving a scanned image with low grey scale contrast, variable background intensity, and noise, locally adaptive binarization methods may improve results (Trier and Taxt 1994). Character and symbol recognition programs may be used at this stage to labels symbols and texts in the image.

The resulting scanned image may be vectorized (described below) or transformed to another kind of

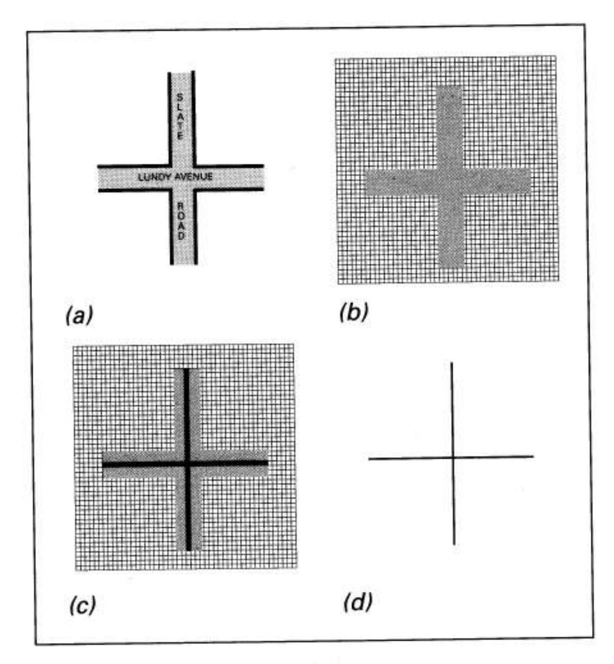


Figure 4.7. Line thinning and skeletization. (*a*) Original map fragment; (*b*) Scanned image; (*c*) Line thinning to capture central pixels; (*d*) Vector representation of road centre lines

raster structure for direct input to the GIS. Matching, scale correction, and alignment may be controlled through known point locations, such as registration or fiducial marks which may be defined directly from the scanned image and processed automatically. Once this is completed the digital map data may be then input to the GIS using an exchange transfer format.

Vectorization Vectorization (raster to vector conversion) is usually undertaken using specialist software which provides algorithms converting arrays of pixels to line data. The process involves threading a line through the pixels of the scanned image using what are known as thinning algorithms. These reduce the pixelated lines to only one pixel wide (shown in Figure 4.7*a–c*). They are then linked from linear or areal units using automatic algorithms which scan and join neighbouring pixels of the same value, or through user-controlled operations.

The resulting data often require much editing to code the individual units and to correct errors in the connectivity of the lines especially if fully automatic vectorization methods are used. Various systems such as Laser-Scans VTRACK and Hitachi CAD-Core tracer systems have been developed which give the operator more control over the unit extraction process (Jackson and Woodsford 1991). The scanned image is displayed on the screen and the operator clicks with the mouse on the starting-point on the feature to be captured. The original system used a projection with two lasers but today the process is carried out by computer software. The computer follows the line until it arrives at a junction, or back to the first coordinate in the case of a closed polygon. Once a line has been scanned, it is 'painted out' of the display. The operator then guides the scanning laser to the next starting-point, and so the process continues. Errors in following the lines may be minimized by the operator defining their shape, for example rectilinear or curvilinear, prior to tracing so different searching algorithms may be used. Where a line is traced incorrectly the operator is able to erase it immediately. These systems have the great advantage that the line scanning is almost instantaneous, and the scanned data are directly digitized in a scale-correct vector (entity) format. The greatest disadvantage is that a great deal of operator control is essential.

An alternative means of converting scanned imagery into a series of points, lines, or polygons is to digitize directly from the image display. These hybrid systems use the raster image essentially as a screen backdrop and the geometry of the various entities is defined by digitizing from the screen using a cursor controlled by a mouse. This method has been used successfully not only with scanned maps but also with digital photographs and satellite images (see Figure 4.3).

Analytical stereoplotters A third type of technology used for capturing digital geographical data is a stereoplotter. This is a photogrammetric instrument used to record the levels and positions of terrain and entities directly from stereo pairs of aerial photographs (taken of the same area but from a slightly different viewing positions). In recent developments, digital stereo images from satellite sensors, video recordings, and digital cameras have been used to generate elevation data using specialized photogrammetric algorithms in image processing systems. The resulting altitude and spatial resolution of the data is variable, but with new satellites being launched giving finer detail this method offers real promise for generating digital elevation maps in the future. New photogrammetric methods are also being developed to generate digital three-dimensional maps. Beers (1995) describes a photographic fisheye recording system that enables a large field of view to be recorded in an image, which may subsequently be used to generate three-dimensional views and coordinates.

Stereoplotters are used extensively in capturing continuous elevation data for digital elevation models and orthophoto maps and they are described in more detail in Chapter 5.

Entering the attribute data Attribute data (sometimes called feature codes) are those properties of a spatial entity that need to be handled in the GIS, but which are not themselves spatial. For example, a road may be captured as a set of contiguous pixels, or as a line entity and represented in the spatial part of the GIS by a certain colour, symbol, or data location (overlay). Information describing the type of road (e.g. motorway or dirt track) may be included in the range of cartographic symbols normally available. If other data about the road such as the width, the type of surface, any specific traffic regulations, the estimated number of vehicles per hour etc., also need to be recorded then these have to be included in the GIS. Although these attribute values and associated identifiers may be attached to graphic entities directly on input, it is not efficient to enter large numbers of complex attributes interactively. The data are therefore either stored separately from the spatial information in the GIS in the case of relational databases, or are input along with spatial description with the object-oriented databases.

Attribute data may come from many different sources such as paper records, existing databases, spreadsheets, etc. They may be input into the GIS database either manually or by importing the data using a standard transfer format such as TXT, CSV, or ASCII. Where relational databases are used an identifier (discussed in more detail in the next section) is included in the attribute record to link the spatial and attribute data together.

Data verification and editing Once the data have been entered it is important to check them for errors—possible inaccuracies, omissions, and other problems—prior to linking the spatial and the attribute data. Figure 4.8 summarizes the complete process of creating a GIS database by hand.

The best way to check for errors in the spatial data is to produce a computer plot or print of the data, preferably on translucent or thin paper, at the same scale as the original. The two maps may then be placed over each other on a light table and compared visually, working systematically from left to right and up and down the map. Missing data, locational errors, and other errors should be clearly marked on the printout. Certain GIS show some topological or identifier errors directly by colour coding the screen image. If the map is a unique drawing, locational errors need only be considered within the map boundary; if the map is one of a series covering a larger area, or the digitized data must link up with map data already in the computer, then the spatial data must also be examined for spatial contiguity across map edges. Certain operations, such as polygon formation, may also indicate errors in the spatial data.

Attribute data may be checked by printing out the files and checking the columns by eye. A better method is to scan the data files with a computer program that can check for gross errors such as text instead of numbers, numbers exceeding a given range, and so on.

Errors that may arise during the capturing of spatial and attribute data (discussed in more detail in Chapter 8) may be grouped as follows:

Spatial data are incomplete or double.
- Incompleteness in the spatial data arises through omissions in the input of points, lines, or cells of manually entered data. In scanned data the omissions are usually in the form of gaps between lines where the raster–vector conversion process has failed to join up all parts of a line. Similarly, the raster–vector conversion of scanned data can lead to the generation of unwanted 'spikes'. Digitized lines may also be input more than once, lines and nodes may not be joined for intersections.

Spatial data are in the wrong place
- Mislocation of spatial data can range from minor placement errors to gross spatial errors. The former are usually the result of careless digitizing; the latter may be the result of origin or scale changes that have somehow occurred during digitizing, possibly as a result of hardware or software faults.

Spatial data are defined using too many coordinate pairs
- As a result of both scanning or digitizing processes lines in the database may be defined using too many points. These may take up exorbitant amounts of storage space.

Spatial data are at the wrong scale
- If all spatial data are at the wrong scale, then this is usually because the digitizing was done at the wrong scale. With scanned data the problems usually arises during the georeferencing process when incorrect values are used.

Spatial data are distorted
- The spatial data may be distorted because the base maps used for digitizing are not scale correct. Most aerial photographs are not scale correct

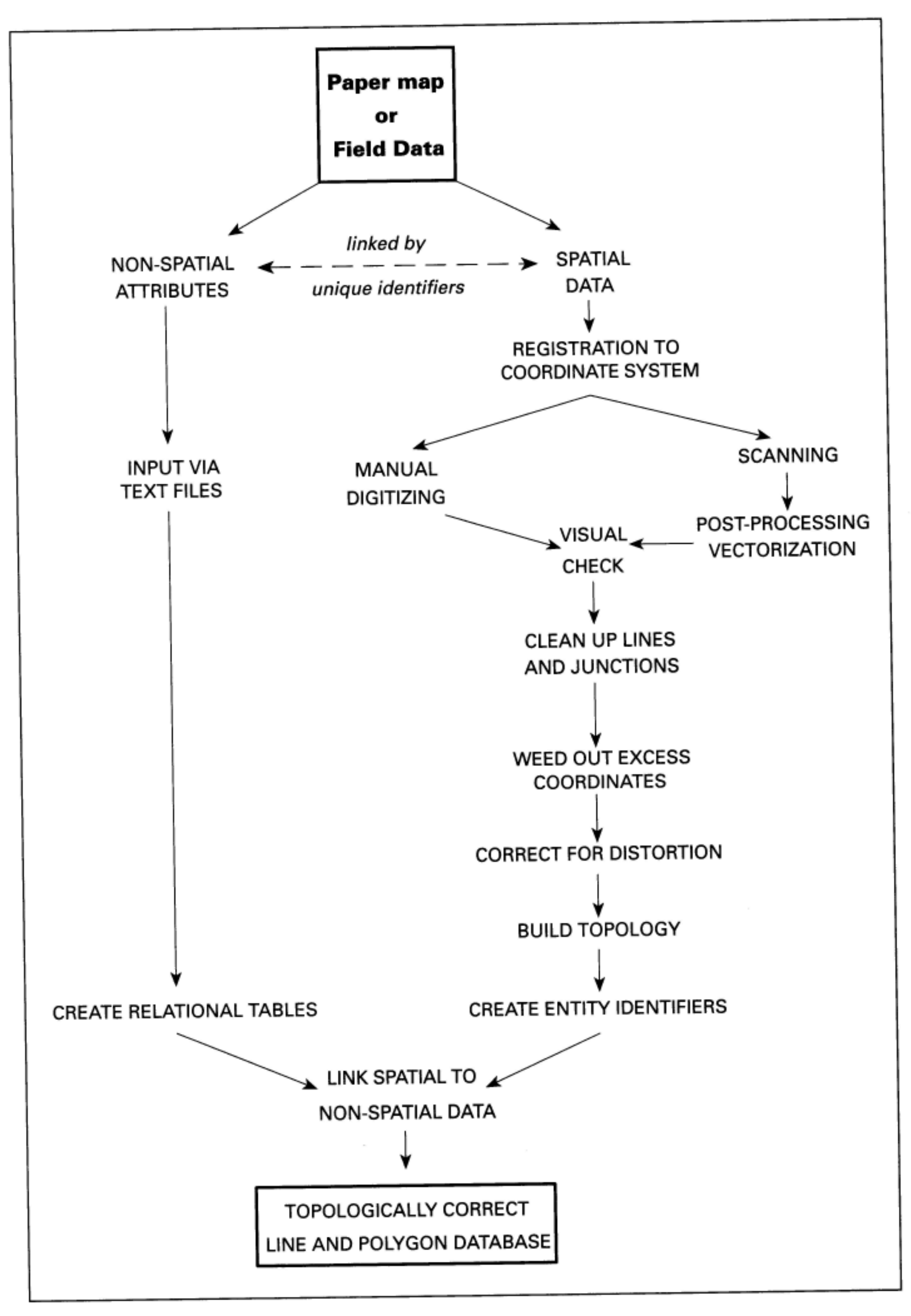


Figure 4.8. Steps in creating a topologically correct vector polygon database

over the whole of the image because of tilt of the aircraft, relief differences, and differences in distance to the lens from objects in different parts of the field. All paper maps suffer from paper stretch which is usually greater in one direction than another. In addition, paper maps and field documents may contain random distortions as a result of having been exposed to rain, sunshine, coffee, beer, and frequent folding. Transformations from one coordinate system to another,

for example, to Universal Transverse Mercator (UTM), may be needed if the coordinate system of the database is different from that used in the input document or image.

These errors need to be addressed through various editing and updating functions supported directly by most GIS. This is a time-consuming, interactive process that can take as long or longer than the data input itself. Data editing is usually undertaken by viewing the portion of the map containing the errors on the computer screen and correcting them through the software using the keyboard, screen cursor controlled by a mouse, or a small digitizer tablet.

Minor locational errors in a vector database may be corrected by moving the spatial entity through the screen cursor or by indicating their position on the digitizer tablet. In some GIS, computer commands may be used directly to move, rotate, erase, insert, stretch or truncate the graphical entities as required. New data may be added through the digitizer or keyboard. Some entity editing operations cannot be used in isolation but must be followed by checks or operations to ensure the coherency of the database. For example, in networks in utility mapping (e.g. telephone lines) editing a single line into two branching lines requires the pointers indicating the flow of signals to be rebuilt. In polygon networks, if a line or part of a line is moved or changed the polygon areas must be recomputed.

Data scaling problems may be overcome by applying simple numerical factors to the data. More complex rotating and translating operations are needed when fitting various data sets together such as a distorted map to an accurate base map. The faulty map should be compared with the base map and a number of points on the original map linked by vectors to what should be their correct positions (Figure 4.9). Mathematical transformations stretch and compress the original map until the linking vectors have shrunk to zero length and the tie points are registered with each other. It is then assumed that all the other points on the original map have been relocated correctly. This process is known as *rubber sheeting* or warping because the original, faulty map is stretched in all directions like an elastic sheet to fit the other. It cannot be used directly on rasterized data because of the rigidity of the fixed grid and the structure of the data; suitable techniques for warping raster structures are described in Chapter 5.

Where excess coordinates define a line these may be removed using 'weeding' algorithms, the best

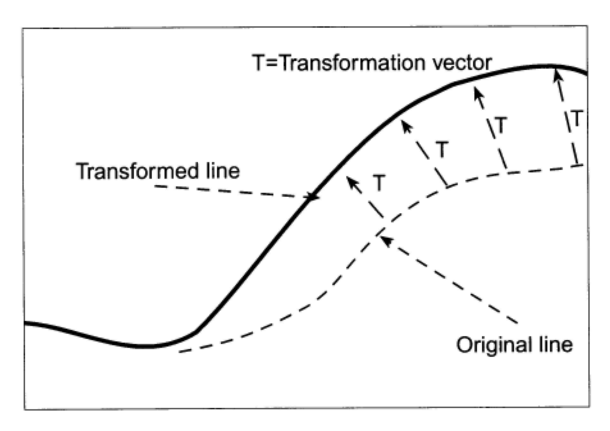


Figure 4.9. Transformation vectors for 'rubber sheeting'

known of which are by Douglas and Peucker (1973) and by Reuman and Witkam (1974). The visual appearance of the resulting set of straight line segments may be 'improved' by using B-splines (see Chapter 5).

Attribute values and spatial errors in raster data must be corrected by changing the value of the faulty cells. This may often be done by a simple command that digitizes the cell followed by inputting the correct attribute value. If large numbers of cells are wrong, the new information could be digitized and simply written over the existing values.

Once the spatial errors have been corrected, the topology of the vector line and polygon networks can be generated as explained in Chapter 3.

Linking spatial and attribute data In GIS with relational database structures the final process in the manual capture of data involves the linking of the attribute and spatial databases through identifiers which are common to the records in both. The road referred to earlier may be given an identifier of *A1* during the data input process; records in the attribute database referring to that road will also need to contain a value that identifies it as *A1*. The links are established for the first time and verified during this data processing stage but are not immutable and are re-established each time the user calls particular parts of the data set.

Identifiers for point and line data are either automatically generated or need to be added manually during the digitizing or scanning/vectorization processes. Polygon identifiers are usually added after the generation of the topology; once the polygons have been formed they may then be given unique identifiers, either by interactive digitizing, or by using 'point-in-polygon' algorithms to transfer identifier codes

from already digitized points or text entities to the surrounding polygon.

The linkage operation provides an ideal chance to verify the quality of both spatial and attribute data. Screening routines can check that every graphic entity receives a single set of attribute data; they can also check that none of the spatial attributes exceed their expected range of values, or that non-sensible combinations of attributes or of attributes and geographical entities occur. All geographical entities not passing the screening test may be flagged so that the operator can quickly and easily repair the errors. Correcting attribute errors requires referring back to the original data to check the information and keying in any missed or wrong values.

The programs that link spatial and attribute data may also be used to check that all links have been properly made. The programs should be written in such a way that they flag only the errors. Incorrect links between spatial and attribute data are usually the result of incorrect identification codes being entered in the spatial data during digitizing or scanning processes, or when establishing polygon topology. Editing the spatial data involves correcting any errors in the polygon topology and the identifier codes given to the data. It is much more difficult to spot errors in attribute data when the values are syntactically good but incorrect.

Data structuring

Following the data capture process the data are now in one of two forms—geometrically correct raster data, or topologically and geometrically correct vector data. Modern GIS will accept both kinds of data but it may be necessary to restructure the spatial data from one form to the other, depending on the application (see Chapters 7 and 8). Vector-to-raster and raster-to-vector conversion have already been discussed earlier in this chapter.

Structuring of the database as a whole is required to optimize the storage requirements and querying performance of the GIS through sensible ordering and data compression techniques as described in Chapter 3. For example, quadtree encoding of data is supported in at least one commercial GIS, and is carried out using basic menu commands. Efficiency within the database is brought about by indexing the attribute and spatial data or through normalization of the RDBMS (again described in Chapter 3).

Data presentation

Once the database is complete, querying and analysis operations may be undertaken, and these are described in detail in Chapters 7 and 8. The results of data retrieval and analyses need to be either in a form that is understandable to a user or that allows data transfer to another computer system. Most GIS include software which support the production of a range of data output options. Digital or *'computer-compatible'* output may be in the form of a file on an optical disk or other storage device that can be subsequently reused or read into another system. Alternatively the data output may involve some form of data transmission over fibre optic or telephone lines such as the Internet.

Analogue or *'people-compatible'* kinds of output are maps, graphs, and tables; the output devices for producing these can be classified into those that produce ephemeral displays on electronic screens and those that produce permanent images on stable base materials such as paper or mylar. The capabilities for producing aesthetically pleasing graphical output have increased in recent years with many vendors developing special mapping tools which are linked to or integrated within the GIS database. These allow detail such as map legend, title, orientation indicator (north arrow or coordinate marks) and a scale or scale bar, as well as colour, symbology, and text to be added easily to the data presentation. They also give choices in displaying the data such as using smooth isolines, or coloured or shaded choropleth maps. The results are maps of high cartographic quality for output on three types of devices: visual display units (VDUs), plotters and printers.

Visual display units

The computer screen or visual display unit is the main tool for displaying the results of GIS analysis to people, especially for Internet users who work on-line.

Electronic computer screens come in a wide range of sizes and use conventional cathode ray tube (CRT) technology or arrays of light emitting diodes (LEDs) to form images on the screen. Within the CRT there are red, green, and blue electron guns which emit beams of light onto coated, light-sensitive particles which make up the screen. When the beam hits one of the particles it emits a coloured light and the image is built up of a grid of thousands of glowing dots. The beams from the CRT are moved across the screen systematically in horizontal and vertical directions to excite the particles in turn and the whole screen is scanned or refreshed 60–80 times a second. The LED displays have triple layers, one for each primary colour.

The emission of light from the CRT to a particular part of the screen is controlled by the GIS software and the graphics adapter on the computer. The software sends a signal to the processor which in turns outputs to the graphics adapter. This latter is a separate card or an addition to the processor board within the computer which controls the signals sent to the CRT and has a small memory capacity for graphics manipulation purposes.

The differences in output resulting from these devices depend on the hardware itself (the VDU and graphics adapter) as well as the GIS display software. They determine resolution of the screen which controls the detail that may be displayed, the number of colours used, and the size and scale of the data shown. Often the displays cannot show all the visual detail of a complex database at full resolution so most systems allow the user to 'zoom in' and display an enlarged part of the database.

Most GIS users will view the results of their analysis as a static map, graph, or table on the screen. Hardcopies of these images may be obtained photographically using special devices or committing them to a temporary file in the computer memory (a screen dump) which may subsequently be printed. Recent advances in computer technology, which enable increasing amounts of data to be stored and viewed, permit more complex screen displays in which a series of images may be shown in rapid succession. This dynamic visualization allows the sensation of motion, either of the viewer or of the environment, to be represented and is particularly useful in GIS modelling applications where changes in the study area are taking place over time. This is a great improvement on previous displays where the values for each successive time step were shown as single static images making it hard for the user to interpret the results. Very few commercial GIS support dynamic visualization capabilities and specialist software is needed to generate this type of display.

The perspective, quasi three-dimensional display or block diagram (otherwise known as *draping*) is a popular method of displaying surficial thematic information in relation to the landforms on which it occurs. The method can be used to provide visual impressions that are impossible to achieve with the usual two-dimensional map format. Examples of perspective displays are given in Plate 1.4 (orthophoto draped over the digital elevation model of the landscape from which it was derived—see Chapter 5), Plates 3.7 and 3.8, and in Figures 5.13, 8.13, and 8.16. Draped images are usually best done in colour because the human eye cannot distinguish sufficient grey scales to make monochrome plots of thematic data fully understandable. The combination of perspective plots and dynamic visualization, as employed in video games, is a powerful means of displaying the results of space-time models in GIS (e.g. Van Deursen and Burrough 1998). Variations within a three-dimensional block of land can be displayed by a range of methods that are marketed by several specialist vendors but fence mapping and cut-away diagrams are common (e.g. Plates 3.3 and 3.4).

Plotters

Plotters are automated draughtsmen: they are output devices for making copies of geographical data on paper or film. They have a paper feeding or holding mechanism, which is usually a roller or flat surface bed, and a drawing 'arm' which houses coloured pens (Figure 4.10). Two-dimensional line images are created by

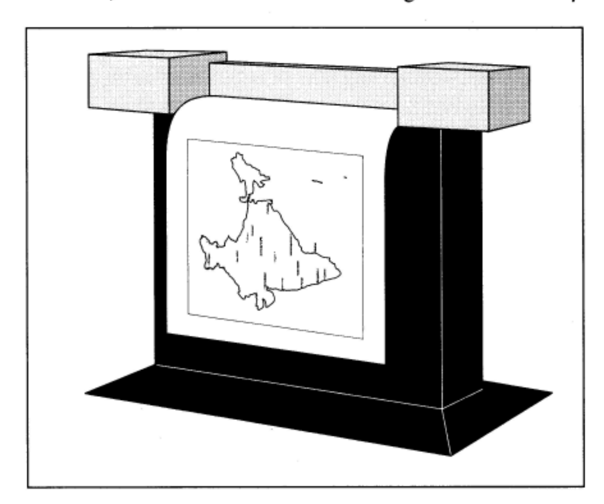


Figure 4.10. Large-format map plotter

moving the drawing device and/or the paper horizontally and vertically (x and y). All information is drawn by a series of line-drawing commands given by the software such as: *move drawing device to point XY, set pen down, move pen to X′Y′, move pen to X″Y″ . . . , raise pen*. The moves to the XY coordinates are given using either absolute or incremental coordinates.

The flexibility and speed of a plotter is determined largely by the amount of preprogrammed information it has for drawing complex shapes such as letters and symbols. 'Smart' plotters will contain preprogrammed, 'hard-wired' character and symbol sets that need only to be referenced by simple computer commands. The quality of the plotted image depends largely on the step-size of the motors controlling the drawing device.

Recently there have been major changes in plotter technology which replace the coloured pens with ink or bubble jet drawing devices, formerly found only in the physically smaller printers (see next section). These have the advantage of being faster and quieter and they allow areal-filled entities as well as linear ones to be drawn quickly and uniformly. They may be used with various media including plain and glossy paper, velum and polyester film. Resolutions of 360 dpi for colour and 720 dpi in monochrome are possible. Laser film writers draw maps on photographic materials to even higher levels of accuracy.

Printers

Printers were originally simple raster output devices such as a line printer or printing terminal which had to use combinations of different keyboard characters to represent different data values. The original computer printers gave images in which the finest resolution was limited to the size of a physical type and so were cartographically unappealing with their coarse grid cell. These have been universally superseded by printers which still build up the images by printing values on a line-by-line basis but which use fine black powder (toner) or coloured inks ejected from a printer head to show the information. These give output of much higher cartographic quality with finer resolutions and up to A0 in size. The resolution is determined by the speed of the step movement and the size of the ink jet with some achieving high-quality output of 720 dpi.

The toner-based printers are known as laser printers and will be familiar to many readers through their use in photocopiers (the xerox process). They work by taking the output from the computer processor and converting it to a laser signal ('on' or 'off') which is temporarily imprinted using a scanning action on to an electrically charged drum. The drum is then passed by a container of toner which is itself electrically charged; those points on the drum where the signal is 'on' attract black powder which sticks to form the image. Paper is passed over this drum and the fine powder is transferred on to it. The paper is next passed between two heated plates which cook the toner so that it gives a permanent copy. Recently, coloured laser printers have become available which work on the same principle but three passes of the laser on to the drum and over the toners are needed for full colour prints.

With inkjet printers the different coloured inks are emitted from fine nozzles in the print head in response to electronic signals from the GIS software. Two main mechanisms have been developed to control the emission. Inkjet printers use electrical methods to transfer ink from cartridges to the paper. Bubblejet printers have a print head which contains a series of nozzles which are approximately one micrometre thick. The ink is released from these nozzles using a deceptively simple mechanism based on the principle that when fluids are heated bubbles are produced. Each nozzle has a heater which when pulsed by an electrical current produces several thousand temperature increases per second. Each of these creates a tiny bubble which exerts pressure within the nozzle, forcing a single, microscopic ink droplet to be ejected. The pressure then drops, a vacuum is created attracting new ink and the process begins all over again.

Data updating

Many geographical data are not inviolate for all time and are subject to change. Few of these changes are so deterministic that they can be performed automatically. For example, political boundaries may change with the whims of parliament, land use and field boundaries may change as the result of reallotment, soil boundaries change as a result of land improvement or degradation. If these changes in the landscape are not included in the database then its credibility and integrity is undermined; updating is therefore needed. The basic operations just described for editing are used to effect this so new data are input, existing information relocated, and new or modified attribute values introduced. When changes to the spatial structuring is needed the topology of the dataset will need to be generated again.

Updating is rather more than just modifying an ageing database; it implies resurvey and processing new information. Some aspects of the earth's surface, such as rock types, change slowly, and important changes are few, so updating remains a small problem. However, there are other kinds of geographical data where it may be more cost-effective to resurvey completely every few years rather than attempt to update old databases. In comparison, updating the attribute data is trivial, provided the one-to-one links between attribute data records and the spatial entities remain unaltered.

Data storage

Building a digital database is a costly and time-consuming process and it is essential that the digital map information is transferred from the local disk memory of the computer to a more permanent storage medium where it can be safely preserved. Some databases for topographic, cadastral, and environmental mapping can be expected to have a useful life of 1–25 years and it is essential to ensure that they are preserved in mint condition.

Digital data are stored on magnetic or optical media however, and the form is variable and reflects the cost of the media, and where and how often the data are used. In the past data were transferred from a computer's hard disk storage (which was relatively expensive to buy) and archived on media such as 0.5 inch wide 9-track magnetic tapes. These were relatively safe, cheap to use, and reliable because the magnetic particles on the tape do not lose their signals over time. They were stored in a safe area and then loaded onto a tape drive mechanism for reading whenever the data were needed. These portable types of media are known as 'removables'. Today rapid changes in hard disk technology have led to more compact and very much cheaper devices being available. More and more data are now being stored on 'non-removables' (fixed storage media) even when the data is used infrequently on the computer.

Removable Storage

Removables are forms of magnetic and optical storage media that may be taken away from the data source computer and used elsewhere. They are used for holding or 'backing up' (for security reasons) data that are currently being used, and for archiving those which are not. The main media available are listed below. The floppy disk and CD ROM are the most familiar products with most desktop machines providing access drives. The other products require special read-write drives to be fitted and are usually used on the main data storage computers.

Magnetic Media are:

- floppy disks—3.5 inch plastic diskettes which contain a flat circular magnetic disk inside and store up to 2.8 megabytes of data but usually 1.4 megabytes; new ZIP drives provide large capacity removable magnetic disks with 100 megabytes to more than 1 gigabyte of data.

- DATs—Digital Audio Tapes which contain magnetic tape within cassettes or cartridges and are in two main sizes of 4 mm and 8 mm; these can store several gigabytes of data,
- 0.5 inch tapes, large spools of magnetic tape which are stored in metal cases, were used most extensively ten years ago though many organizations still use them today, and store about 520 megabytes of data.

Optical media are:

- plastic resin disks which allow a laser to encrypt digital data on to it and their most familiar form is used to record music.
- optical disks (synonymously called CDs—compact disks) which are approximately 13 cm (5.25 in) in diameter, are able to store 675 megabytes, and are of a variety of forms which are distinguished by their ability to write data to the disk:
  - *CD ROMS* (Read Only Memory) act as pure storage media and many data and software are distributed by them;
  - *WORM Optical Disks* (Write Once Read Many) allow the user to copy the data to the disk only once, after which time the data can only be read from the disk;
  - *Rewritable Optical Disks* are usually held within a plastic housing and allow data to be copied to and from the disk many times—they are effectively large storage floppy disks;
  - laser disks are large 12 in disks and are usually ROMS.

There are essentially two types of operation used to commit the data to storage onto these media. *Archiving* is the transfer of the data from the computer onto the media with the specialist software or the hardware drives themselves performing compression operations to maximize the amount of information stored. This data encryption requires the information to be 're-stored' prior to using it in the GIS again. The second method, known as *Mass Storage Systems* (MSS), is now widely used and differs from archiving systems in that the stored data are written in the same format as the original. The data may therefore be accessed directly from the storage media and used in the GIS. Rewritable optical disks are being used increasingly in these operations and some computers have fittings which are able to hold many of these; they are known as optical juke boxes. The computer is then able to call on any one of these devices to provide data.

NON-REMOVABLE STORAGE AND NETWORKS

The falling price of hard disks per megabyte of data has meant that economies in storage costs, formerly made using removables, no longer apply. Gigabytes of data may now be stored on a single computer hard disk. The main problem associated with this is sharing the data; users will have to either use only the computer on which they are stored or make copies of the data to put on their own. This is obviously inconvenient and leads to a large duplication in data storage, and difficulties in knowing which is the most current version of the database. Where there are a number of users wanting to access the same data (and software), organizations have used computer networks to minimize these problems.

There are two main types of network—Local Area Networks (LANs) and Wide Area Networks (WANs). LANs are where computers are linked together at one site; WANs are networks which link geographically remote sites. The computers are joined using either wired (copper-based or fibre optic cable) or wireless (radio, microwave, laser, and infrared) links. Wireless links support point-to-point connections and are used mostly with WANS, whereas wired links are used to physically join the computers together in LANs, and where feasible WANs.

The configuration of computers on a network may be described using the following types:

- Peer-to-peer—where two computers of equal power are joined for the sharing of files. This type of network is little used in GIS environments because transferring the large data files involved seriously affects the performance of the computers.
- Client-server—there are one or more dedicated, powerful computers (servers) which are used to store the data and software. The computers linked to the server (the clients) have their own processing power but access data and software from the central store. The server manages file transfer and storage, central printing, and, with some systems, supports database querying and manipulation tasks.
- Central processing systems—these are principally associated with main-frame and mini networks and consist of an extremely powerful, central processing computer which stores all the data and software, and a series of dumb terminals through which the user accesses them. All the processing is done by the main computer.

Networks used with GIS are of types 2 and 3 with a general move these days towards the client–server approach. This provides efficient data transfer and printing capabilities although the GIS software is usually installed on the clients because of its size. Network based storage is not without its problems. The principal ones are associated with multiple accessing of the data and ensuring the latest version is always made available. Problems are encountered if more than one user is updating or using a version of a database at the same time. There are also problems associated with giving a large number of people access and the ability to change a very valuable data resource. Access and usage limits therefore have to be imposed.

Questions

1. Explain how you would assemble the geographical data needed for the following applications.
   - The location of fast food restaurants
   - The incidence of landslides in mountainous terrain
   - The dispersion of pollutants in groundwater
   - An emergency unit (police, fire, ambulance)
   - A tourist information system
   - The monitoring of vegetation change in upland areas
   - The monitoring of movement of airborne pollutants, such as the ^{137}Cs deposited by rain from the Chernobyl accident in 1986.

2. What steps would you take to limit the introduction of errors in (*a*) the digitizing and (*b*) the scanning of spatial data?

3. Look at the computers you have near to you. What devices and media are used for storing or backing their data? How would you improve the safety factor for this?

4. What are the most important aspects of a map for communicating information? What makes a good map? How should you choose colours and grey scales for displaying data?

Suggestions for further reading

BURROUGH, P. A., and MASSER, F. I. (1998). *European Geographic Information Infrastructures: Opportunities and Pitfalls.* Taylor & Francis, London.

GUPTILL, S. C. (1991). Spatial data exchange and standardization. In D. J. MAGUIRE, M. F. GOODCHILD, and D. W. RHIND (eds.), *Geographical Information Systems, i: Principles.* Longman Scientific and Technical, Harlow, Essex, pp. 515–30.

KENNEDY, M. (1996). *The Global Positioning System and GIS.* Ann Arbor Press, Inc. Ann Arbor, 268 pp.

MALING, D. H. (1973). *Coordinate Systems and Map Projections.* George Phillip, London.

RHIND, D. W. (1992). Data access, charging, and copyright and their implications for geographical information systems. *International Journal of Geographical Information Systems*, 6: 13–30.

FIVE

Creating Continuous Surfaces from Point Data

This chapter explains methods of creating discretized, continuous surfaces for mapping the variation of attributes over space. Data sources are commonly observations at sparsely distributed point samples such as groundwater wells, soil profiles, meteorological stations or presence/absence data for vegetation or counts of animals, people or marketing outlets for basic spatial units such as grids or administrative areas. The results are usually interpolated to regular grids and can be displayed as colour or grey scale maps or by contour lines. The chapter describes spatial sampling strategies and methods of spatial prediction including global methods of classification and regression and local deterministic interpolation methods such as Thiessen polygons, inverse distance weighting and thin-plate splines. Each method is illustrated by worked examples that provide a basic introduction to the mathematics and demonstrate when the method is applicable. Digital elevation models and digital orthophoto maps are examined as special cases of continuous surfaces.

Interpolation: what it is and why it is necessary

Interpolation is the procedure of predicting the value of attributes at unsampled sites from measurements made at point locations within the same area or region. Predicting the value of an attribute at sites outside the area covered by existing observations is called *extrapolation*.

Interpolation is used to convert data from point observations to continuous fields so that the spatial patterns sampled by these measurements can be compared with the spatial patterns of other spatial entities. Interpolation is necessary,

(*a*) when the discretized surface has a different level of resolution, cell size or orientation from that required, or

(*b*) when a continuous surface is represented by a data model that is different from required, or

(*c*) when the data we have do not cover the domain of interest completely (i.e. they are samples).

Examples of (*a*) are the conversion of scanned images (documents, aerial photographs, or remotely sensed images) from one gridded tessellation with a given size and/or orientation to another. This procedure is known generally as *convolution*.

Examples of (*b*) are the transformation of a continuous surface from one kind of tessellation to another (e.g. TIN to raster or raster to TIN or vector polygon to raster).

Examples of (*c*) are the conversion of data from sets of sample points to a discretized, continuous surface. We must distinguish situations with *dense* sampling networks from those with *sparse* sampling networks, or data collected along widely spaced transects. Dense sampling networks are common when creating hypsometric surfaces to represent variations in the elevation of the land surface (*Digital Elevation Models—DEM*; also known as *Digital Terrain Models—DTM*) from aerial photographs and satellite imagery where the source data are cheap and the attribute can be observed directly. Sparse sampling networks are often imposed by the costs of borings, laboratory analyses, and field surveys—most often, the spatial variation of the attribute of interest cannot be seen and must be derived indirectly. The many uses of continuous surfaces in spatial modelling are explained in Chapter 8.

The continuous surfaces obtained from interpolation can be used as map overlays in a GIS or displayed in their own right. The surfaces can be represented by the data models of contour lines (*isopleths*), discrete regular grids, or by irregular tiles. The data structures used are regular grids (raster), contour lines, or triangular irregular networks (TINs). Because the interpolated surfaces vary continuously over space, regular gridded interpolations must be represented by a data structure in which each grid cell can take a different value: therefore space-saving raster data structures such as run-length codes and quadtrees cannot easily be used. The original data may be collected as point samples distributed regularly or irregularly in space (and/or time), or they can be taken from already gridded surfaces such as remotely sensed images and images from document scanners. Attributes predicted by interpolation are usually expressed by the same data type as those measured, but some interpolation methods provide means to estimate *indicator functions* that show a *probability* that a given value is exceeded, or that a given class may occur.

The visualization of continuous surfaces

Continuous surfaces are usually represented by images or lines. Image methods include regular and irregular grids and tessellations in which the variations of the value of the mapped attribute is indicated by graded zones of colour or grey levels. 'Two and a half D' representation is achieved by 'draping' the attribute surface over a continuous surface that represents the topography of the land (e.g. Plates 1.4, 4.3, or Figure 5.13).

Line representation includes *isolines* (lines of equal value), vertical slices (profiles) (Plate 3.3), and critical lines, such as ridges, stream courses, shorelines, and breaks of slope. Line methods and image methods can be combined to enhance perception.

Although interpolated surfaces show variation along three data axes, namely the x and y coordinates and the axis of the interpolated attribute, they are usually not considered as 3D representations. The term 'three-dimensional' is usually (and properly) reserved for situations in which an attribute varies continuously through a 3D spatial frame of reference. True 3D and 4D representation and visualization requires special software not usually found in standard GIS toolkits (Nichols *et al.* 1992). Continuous volumes are usually represented by fence diagrams (slices through a 3D image—Plates 3.3 and 3.4) or by 3D display of a crisped image of a zone of concentration, such as an ore body or a pollution plume.

The rationale behind interpolation

The rationale behind spatial interpolation and extrapolation is the very common observation that, on average, values at points close together in space are more likely to be similar than points further apart. In general, two observation points a few metres apart are more likely to have the same altitude than points on two hills some kilometres apart. However, *classification* is also a popular method for predicting values at unsampled locations from estimates of map unit means or the central concepts of taxonomic classes, irrespective of any spatial association between the measured values within the classes. When mean attribute values for 'homogeneous' classes or pieces of land have been computed, for example for a given map unit on a geological, soil, land cover, or vegetation map, all information on the short-range variation of the attributes has been lost. Most methods of interpolation attempt to use this local information to provide a more complete description of the way an attribute varies within that area. If this variation can be captured successfully, we may expect that our estimates of the value of any given attribute at unvisited sites will be better than those obtained from class averages alone. The maps so constructed should result in smaller errors when used for subsequent overlay analyses and quantitative modelling in the GIS.

Data sources for interpolation

Sources of data for continuous surfaces include:

- Stereo aerial photos or overlapping satellite images using photogrammetry.
- Scanners in satellites or aeroplanes and document scanners
- Point samples of attributes measured directly or indirectly in the field on random, structured, or linear sampling patterns, such as regular transects or digitized contours.
- Digitized polygon/choropleth maps.

Many data for interpolation come from *sampling* a complex pattern of variation at relatively few points. These measurements are often known as 'hard data'. When data are sparse, it is very useful to have information on the physical processes or phenomena that may have caused the pattern which are known as 'soft information', and which can assist interpolation. In many cases, however, the physical process is unknown and we must make do with various kinds of assumptions about the nature of the spatial variation of the attribute in question. These include assumptions about the smoothness of the variation sampled, and statistical assumptions concerning the probability distribution and statistical stationarity (see next chapter).

Spatial sampling

The location of sample points can be critical for subsequent analysis. Ideally, for mapping, samples should be located evenly over the area. A completely regular sampling network can be biased, however, if it coincides in frequency with a regular pattern in the landscape, such as regularly spaced drains or trees, and for this reason statisticians have preferred to have some kind of random sampling for computing unbiased means and variances. Completely random location of sample points has several drawbacks too. First, every point has to be located separately, whereas a regular grid needs only the location of the origin, the orientation, and spacing to fix the position of every point. This is much easier in wooded or difficult terrain, even with GPS. Second, complete randomization can lead to an uneven distribution of points unless very many points can be measured, which is usually prohibited by costs.

Figure 5.1 presents the main options available. A good compromise between random and regular sampling is given by stratified random sampling where individual points are located at random within regularly laid out blocks or strata. Cluster (or nested) sampling can be used to examine spatial variation at several different scales. Regular transect sampling is often used

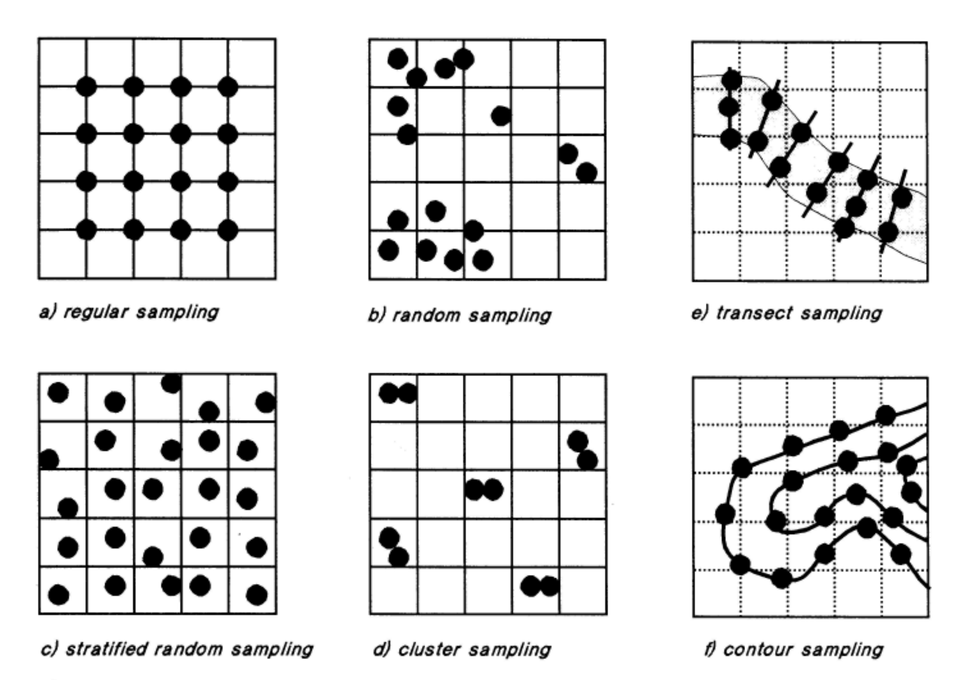


Figure 5.1. Different kinds of sampling net used to collect spatial data from point locations

to survey profiles of rivers, beaches, and hillsides, and digitizing contour lines is a common method of sampling printed maps to make digital elevation models.

THE SUPPORT

The *support* is the technical name used in geostatistics for the area or volume of the physical sample on which the measurement is made. If the sample is a kilogram of soil taken from a soil profile pit, then the support would be approximately 10 × 10 cm in area and about 5 cm thick. If the sample is a litre of groundwater extracted from a sampling tube then the support has the dimensions of the column. Because analytical laboratory methods usually homogenize the samples by grinding or mixing, all internal structure or variation is lost. Therefore the measurement refers to the area or volume of material sampled and not to a larger or smaller area or volume.

When data collected on a given support are used to predict values of the same attributes at unsampled locations then the predictions refer to locations that also have that support, unless procedures such as bulking or spatial averaging are used to relate the observations to larger areas or volumes. These procedures are known collectively as 'upscaling' procedures. The simplest upscaling procedure is to collect a bulk sample consisting of several small samples taken within a defined area around the geometrically located sampling 'point' and to homogenize this bulked sample before laboratory analysis. For example if 10 sub-samples were collected and mixed within an area of 10 × 10 m the support would then be a square of the same dimensions.

Enlarging the support by bulked sampling is sensible when the short-range spatial variation of the attribute in question is so large that long-range variations cannot be clearly distinguished. It is also useful when data collected by different methods need to be combined, such as soil or water information and information collected by remote sensors. Most remote sensing scanners collect data which are recorded as single grid values (pixel values) and pixels can vary in size from a few centimetres to several kilometres. The numbers recorded are area-weighted averages of the radiation received, so the pixel area defines the size of the support. Ground measurements, however, are usually made at much smaller locations and will detect variations within the pixels, unless suitably bulked. If the support sizes of both sets of observations are not matched it may be difficult to combine the data sets for modelling or spatial analysis (Burrough 1991*a*).

In demographic studies the support is often an irregularly shaped area determined by a census district or postcode area. These vary in size and shape (and sometimes over time) so they provide a difficult basis for

interpolation. Frequently the data collected are not measurements of a single attribute such as elevation or barometric pressure, but *counts* of people or socio-economic indicators expressed on a nominal or ordinal scale. For these kinds of data interpolation may serve to bring data from different kinds of support (post codes and census districts) to a common base.

TERMINOLOGY

Throughout this book we shall use the following terminology for data sampled at point locations. The attribute value at a data point is denoted by $z(x_i)$, where the subscript i indicates one of a number n of possible measurements that are geographically referenced to the coordinates $\mathbf{x}$ of any convenient cartesian grid. A predicted value at an unsampled location is indicated by $\hat{z}(x_0)$.

Exact and inexact interpolators An interpolation method that predicts a value of an attribute at a sample point which is identical to that measured is called *an exact interpolator*. This is the ideal situation, because it is only at the data points that we have direct knowledge of the attribute in question. All other methods are *inexact interpolators*. The statistics of the differences (absolute and squared) between measured and predicted values at data points $\hat{z}(x_i) - z(x_i)$ are often used as an indicator of the quality of an inexact interpolation method.

Methods for interpolation

The methods of interpolation discussed in this chapter include:

- *Global methods*
  classification using external information
  trend surfaces on geometric coordinates
  regression models on surrogate attributes
  methods of spectral analysis
- *Local deterministic methods*
  Thiessen polygons and pycnophylactic methods
  linear and inverse distance weighting
  thin plate splines.

All these methods are relatively straightforward, requiring only an understanding of deterministic or simple statistical methods. These methods are frequently included in commercial GIS, though the methods of operation are seldom given.

Geostatistical methods using methods of spatial autocorrelation, known as 'kriging' require an understanding of the principles of statistical spatial autocorrelation. These methods are used when the variation of an attribute is so irregular, and the density of samples is such, that simple methods of interpolation may give unreliable predictions. Geostatistical methods provide probabalistic estimates of the quality of the interpolation. Also, they enable predictions to be made for blocks of land greater than the support. In addition, geostatistical methods permit the interpolation of indicator functions and can incorporate soft data to guide interpolation, thereby increasing the precision of the results. Some GIS include simple geostatistical methods but usually these are limited in scope and it is better to export the data to a specialized geostatistical package. Because of the theoretical background, geostatistical methods of interpolation are described separately in Chapter 6.

THE EXAMPLE DATA SET

In order to provide a fair comparison of all the different interpolation methods used in this chapter and Chapter 6 to interpolate from sparsely sampled populations, we use a single data set (see Figure 5.2*a* and Appendix 3). Here and in Chapter 6 we use 98 observations taken from a larger set of 155 soil samples from the top 0–20 cm of alluvial soils in a 5 × 2 km part of a larger study area on the floodplain of the River Meuse (Maas) near the village of Stein in the south of the Netherlands. All 'point' data refer to a support of 10 × 10 m, the area within which bulked samples were collected using a stratified random sampling scheme. Elevation of sample sites and their distance from the river were also recorded. Samples were chemically analysed for concentration of heavy metals (cadmium, zinc, lead, and mercury); here we use the reported zinc content in ppm for comparative illustration of all methods of interpolation.

Global interpolation

Methods of interpolation can be divided into two groups, called *global* and *local* interpolators. Global interpolators use all available data to provide predictions for the whole area of interest, while local interpolators operate within a small zone around the point being interpolated to ensure that estimates are made only with data from locations in the immediate neighbourhood, and fitting is as good as possible.

Global interpolators are mostly used not for direct interpolation but for examining, and possibly removing, the effects of global variations as caused by major trends or the presence of various classes of land that may indicate areas that have different average values. Once the global effects have been take care of, the *residuals* from the global variations can be interpolated locally.

Global methods are usually simple to compute and are often based on standard statistical ideas of variance analysis and regression. *Classification methods* use easily available soft information (such as soil types or administrative areas) to divide the area into regions that can be characterized by the statistical *moments* (mean, variance) of the attributes measured at the locations within those regions.

Regression methods explore a possible functional relation between attributes that are easy to measure and the attribute to be predicted. These methods can be based only on the geographical coordinates of the sample points (in which case the method is called *trend surface analysis*), or on some relationship with one or more spatially variable attributes (in which case the empirical regression model is often called a *transfer function*).

Methods such as Fourier transform and wavelets are also used for interpolation (particularly in the analysis of remotely sensed imagery—Lillesand and Kiefer 1987, Buiten and Clevers 1993), but as they require large amounts of data at many different levels of resolution they are not considered here.

Global prediction using classification models

When spatial data are sparse it is sometimes convenient to assume that the observations are taken from a statistically stationary population (i.e. the mean and variance of the data are independent of both the location and the size of the support). Alternatively, we can decide that these observations sample some coordinated spatial change. If we opt for the former, we automatically select a classificatory approach to spatial prediction, implying that the spatial structure of variation is determined by these externally defined spatial units. If we opt for classification, we can compute our predictions by using well-known, standard analysis of variance (ANOVA) methods.

Classification by homogeneous polygons assumes that within-unit variation is smaller than that between units; the most important changes take place at boundaries. This conceptual model is commonly used in soil and landscape mapping to define 'homogeneous' soil units, landscape units, ecotopes, etc. where 'objects' like soil mapping units, river terraces, catchment areas, hillsides, breaks of slope, etc. have been recognized as useful features for carrying information about other aspects of the landscape.

The simplest statistical model is the ANOVA model:

$$z(x_0) = \mu + \alpha_k + \varepsilon \qquad 5.1$$

where z is the value of the attribute at location x_0, μ is the general mean of z over the domain of interest, α_k is the deviation between μ and the mean of unit k, and ε is residual (pooled within-unit) error, sometimes known as *noise*.

This model assumes that for each class k the attribute values are normally distributed; ideally, each contains a distinct mode. The mean attribute value per class k is $\mu + \alpha_k$, which is estimated from a set of independent samples that are assumed to be spatially

Box 5.1. BASIC PRINCIPLES OF MEANS AND VARIANCE

The meaning of the mean and variance of a population and a sample

Not all attributes can be measured exactly; there is a natural variation of values around the *mean* or average. In many cases this variation can be described by the *Bell curve*, or normal distribution, whose parameters are the mean μ and standard deviation σ. The width of the normal distribution is given by σ; 65 per cent of all values of a normally distributed population fall within $\mu \pm \sigma$, 95 per cent fall within $\mu \pm 2\sigma$ and 99 per cent fall within $\mu \pm 3\sigma$. Therefore, the larger the standard deviation of a population, the lower the *precision* that can be associated with the value of any sample of that population.

In many cases our population of geographical entities can be regarded as almost infinite (think of all possible soil profiles, ground water observation wells and even fast-food restaurants) so the data collected are only a *sample*. We estimate μ and σ by computing the sample mean m and sample standard deviation s from n observations as:

$$m = 1/n \sum_{i}^{n} z_i$$

and

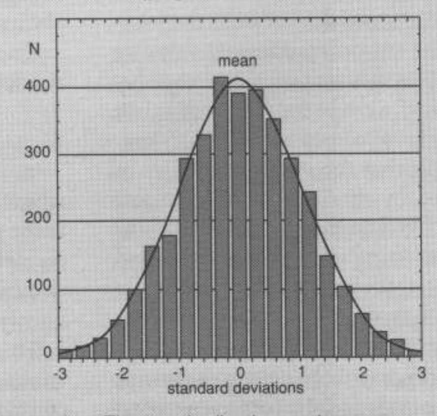
$$s = \left[1/n \sum_{i}^{n} (z_i - m)^2 \right]^{0.5}$$

Obviously it is useful to know s and to ensure that it is as small as possible to obtain the best precision.

The square of the standard deviation is called the *variance* σ^2. The variance is a very useful quantity because variances computed from different sources of variation can be added together to combine estimates of uncertainty from different sources. So, if we have a map with p_k polygons, containing n_i observations per polygon we can compute three variances, namely σ_T^2, the total variance of all observations which is made up of σ_B^2 the *between class variance*, and σ_W^2 the *within class variance* (pooled over all classes). These variances are estimated by our sample which computes:

$$s_T^2 = s_B^2 + s_W^2$$

The value of dividing the data according to the p_k polygons is indicated by the statistical *significance* of the variance ratio s_W^2/s_B^2 for the appropriate *degrees of freedom*. The



The standardized normal curve fitted to experimental data

degrees of freedom are given by $p_k - 1$ for s_B^2 and by $n - p_k$ for s_W^2 and can be found in statistical tables of F.

The magnitude of the variation 'explained' by the soft data classification into p_k polygons is given by s_B^2/s_T^2, which in regression is called R^2. The larger R^2 the more of the variation is explained by the p_k classes, and the more likely that the mean values of each class m_k will be a good global predictor.

The variation within each polygon can be explored by subtracting the data values z_i from the polygon mean m_k in which they are situated.

Table 5.1. One-way analysis of variance of n observations spread over k classes.

Source	Sums of squares	Degrees of freedom	Mean square	Variance ratio
Between	SS_b	$k - 1$	MS_b	MS_b/MS_w
Within	SS_w	$n - k$	MS_w	
Total	SS_t	$n - 1$	MS_t	

independent. The mean within-class variance is given by ε, and is assumed to be the same for all classes.

The relative variance (σ_w^2/σ_t^2), where σ_w^2 is the pooled within-class variance and σ_t^2 the total sample variance, is a measure of the goodness of the classification. Both can be estimated when all map units contain more than one point sample. The lower the relative variance, the better the classification. The analysis of variance is set out in Table 5.1. The significance of the success of the classification can be tested with the usual statistical F-test on the variance ratio with degrees of freedom $m, n - k$.

Assumptions

This approach makes the following assumptions about the spatial variation

- variations in the value of *z within* the map units are random and not spatially contiguous
- all mapping units have the same within class variance (noise) which is uniform within the polygons
- all attributes are normally distributed
- all spatial change takes place at boundaries, which are sharp not gradual.

Note that these assumptions need not necessarily hold. Some map units or individual occurrences of a given map unit might be internally more variable than others. The assumption of within-class variation being random implies that differences cannot be mapped out at larger mapping scales, which is usually not true: soil maps are made over a wide range of nested scales, with differences being seen at all scales. Data are not necessarily normally distributed and may be distributed lognormally, rectangularly, hyperbolically, or in other ways. In some cases, normalization by computing natural logarithms, logit transformations, or other suitable transformations may be necessary (Box 5.2). When transformation is necessary, it is more sensible to treat each map unit, or indeed each delineation of each map unit, as a separate entity and to compute individual means and standard deviations if there are sufficient data.

Spatial prediction of zinc levels in the topsoil by the ANOVA method will be illustrated using soft information from a flooding frequency map with three classes: 1 frequent flooding (annual), 2 flooding every 2–5 years, and 3 flooding less than every 5 years. As the zinc is carried to the site by polluted river sediment a reasonable hypothesis is that frequently flooded sites will have greater concentrations of zinc.

Figure 5.3 shows that, as a group, the zinc measurements do not appear to be normally distributed with a single mode. They could be log-normally distributed or they could be a multimodal distribution with different means and variances. To see the effect of

BOX 5.2. COMMON DATA TRANSFORMATIONS

Transformations commonly used to bring data to a normal distribution

1. *Logarithms. When data are very strongly positively skewed, with a small number of values that are much larger than the mean or mode, the data can be normalized by computing logarithms to base 10 or base e. There must be no zero or negative values in the data, so a small constant can be added to make all data greater than zero, i.e. the natural logarithm is given by:*

$$U = ln(A + c)$$

where U is the transformed variable, A is the original variable, and c is a small constant to ensure all values exceed zero.

2. *Logit. The logit transformation is used to spread out distributions of data on proportions (range 0–1 or 0–100 as per cent) so that concentrations at the ends of the range are avoided. The logit transformation is:*

$$U = ln(p/q)$$

where U is the transformed variable, p is an observed proportion, and q = 1 – p. Zero and unity values of p should be avoided by adding or subtracting a small constant, as before.

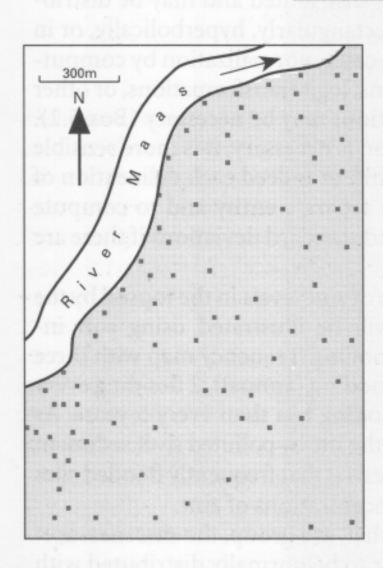
3. *Square root transformation. Moderate skewness can be removed by computing:*

$$U = (A)^{0.5}$$

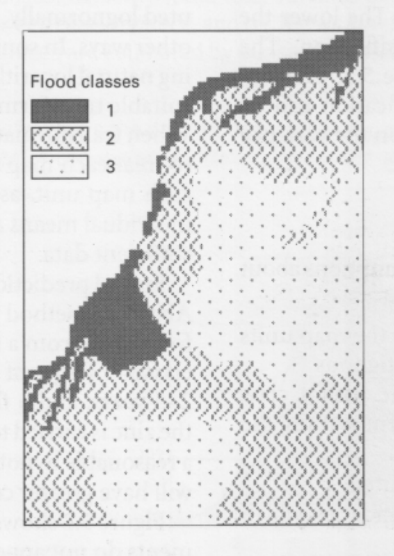
4. *Angular transformations. The arcsine transformation is used for proportional counts to spread the distribution near the ends of the range. If the proportion is p then:*

$$U = sin^{-1}(p)^{0.5}$$

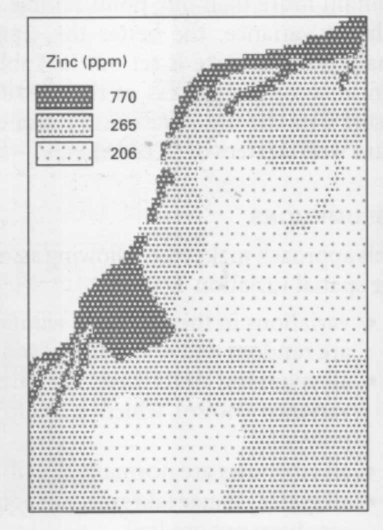
i.e. U is the angle whose sine is $p^{0.5}$



(a) Data points

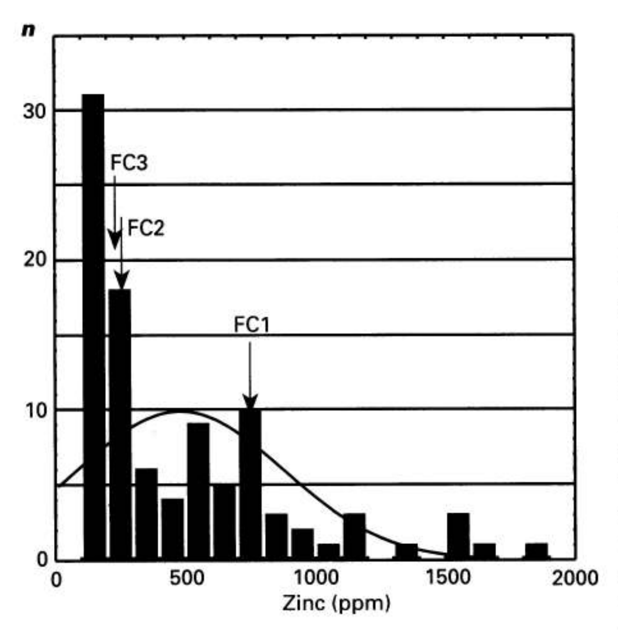


(b) Flood frequency



(c) Zinc levels predicted by flood frequency

Figure 5.2. (*a*) Location of sites of example data set, (*b*) flood frequency classes, and (*c*) zinc levels per flood frequency class



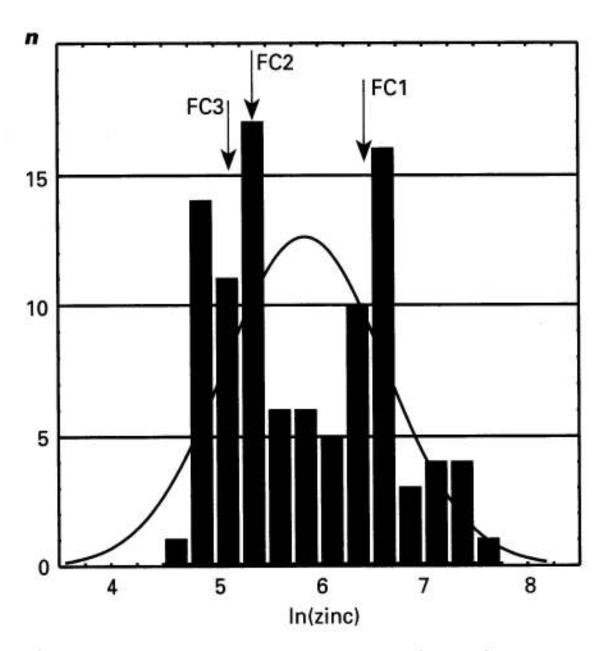


Figure 5.3. (Top) histogram of zinc levels; (bottom) histogram of ln(zinc)

non-normality the analysis of variance is carried out both for untransformed data and data transformed to natural logarithms.

Untransformed zinc levels The class means and standard deviations are (units ppm zinc):

	Mean	Standard deviation
1	769.76	423.17
2	264.97	176.62
3	205.77	105.32

Figures 5.2*b*,*c* show the flood frequency map and the derived map of the zinc content of the soil obtained by assigning the mean value of each class as computed from the data points located in the respective flood frequency classes. Because the maps show only average values per class, the maps of untransformed and transformed zinc levels look the same. The conclusion to be drawn from this analysis is that only those areas that are annually flooded have elevated levels of zinc in the soil. Clearly, the flood frequency map distinguishes the zinc content of the soil of the most frequently flooded areas from those less frequently inundated, but as Figure 5.3 shows, there is a large overlap between classes 2 and 3. The standard deviations per class are large because of the non-normal distribution and so the 95% confidence intervals span zero, which is not sensible (−76.6–1616.0 ppm zinc).

Lognormal transformation The class means, standard deviations and back-transformed means are:

	Mean(ln)	Standard deviation(ln)	Mean (exponent)
1	6.484	0.609	654.58
2	5.421	0.531	226.11
3	5.239	0.415	188.48

The analysis of variance is given in Table 5.3.

These results show that the log transformation has greatly increased the variance ratio, and hence the quality of the classification. The 95 per cent confidence intervals on the log data do not span zero, so they are sensible but large (193.6–2212.8 ppm zinc). A student's t test on the logarithmic map unit means demonstrates that no statistical significance can be attributed to the differences between map units 2 and 3. Together with the histogram views of the data (Figure 5.3), the results confirm the large difference between flood frequency class 1 and the other two classes, which overlap considerably. The results suggest that we could usefully pool the data from flood classes 2 and 3 thereby modifying our soft information to two classes (frequently flooded versus infrequently flooded) with different means *and* variances. Simplifying the classification implies that it is the frequent (annual) floods that are most responsible for depositing polluted sediments (cf. Middelkoop 1997).

Table 5.2. Analysis of variance of zinc by flood classes.

Source	Degrees of freedom	Mean square	Variance ratio	Relative variance
Between	2	3206979	33.8	0.60
Within	95	94880		
Total	97	159048		

Table 5.3. Analysis of variance of zinc transformed to natural logarithms

Source	Degrees of freedom	Mean square	Variance ratio	Relative error
Between	2	14.55	46.60	0.52
Within	95	0.312		
Total	97	0.6056		

Global interpolation using trend surfaces

When variation in an attribute occurs continuously over a landscape it may be possible to model it by a smooth mathematical surface. There are several ways of doing this: all of them fit some form of polynomial equation to the observations at the data points so that values at unsampled locations can be computed from their coordinates.

The simplest way to model long-range spatial variation is by a multiple regression of attribute values versus geographical location. The idea is to fit a polynomial line or surface, depending on whether our data are in one or two dimensions, by least squares through the data points thereby minimizing the sum of squares for $\hat{z}(x_i) - z(x_i)$. It is assumed that the spatial coordinates (x,y) are the independent variables, and that z, the attribute of interest and the dependent variable is normally distributed. Also, it is assumed that the regression errors are independent of location, which is often not the case.

As a simple example, consider the value of an environmental attribute z that has been measured along a transect at points x_1, x_2, x_n. If, apart from minor variation, the value of z increases linearly with location, **x**, its long range variation can be approximated by the regression model

$$z(x) = b_0 + b_1\mathbf{x} + \varepsilon \qquad 5.2$$

where b_0 and b_1 are the polynomial coefficients known respectively as the intercept and the slope in simple regression. The residual ε (the noise) is assumed to be normally distributed and independent of the **x** values.

In many circumstances z is not a linear function of **x** but may vary in a more complicated way. Quadratic or still higher order polynomial models such as

$$z(x) = b_0 + b_1\mathbf{x} + b_2\mathbf{x}^2 + \varepsilon \qquad 5.3$$

can be used. By increasing the number of terms it is possible to fit any set of points by a complicated curve exactly, thereby reducing ε to zero.

In two dimensions the polynomials derived by *multiple regression* on x and y coordinates are surfaces of the form

$$f\{(x,y)\} = \sum_{r+s\leq p} (b_{rs} \cdot x^r \cdot y^s) \qquad 5.4$$

of which the first three are:

b_0	flat	5.5
$b_0 + b_1 \cdot x + b_2 \cdot y$	linear	5.6
$b_0 + b_1 \cdot x + b_2 \cdot y + b_3 \cdot x^2 + b_4 \cdot xy + b_5 \cdot y^2$	quadratic	5.7

The integer p is the order of the trend surface. There are $P = (p + 1)(p + 2)/2$ coefficients that are normally chosen to minimize

$$\sum_{i=1}^{n}\{z(\mathbf{x}_i) - f(\mathbf{x}_i)\}^2 \qquad 5.8$$

where $\mathbf{x}$ is the vector notation for (x,y). So a horizontal plane is zero order, an inclined plane is first order, a quadratic surface is second order, and a cubic surface with 10 parameters is third order.

Finding the b_i coefficients is a standard problem in multiple regression, so the computations are easy with standard statistical packages. Trend surfaces can be displayed by estimating the value of $z(\mathbf{x})$ at all points on a regular grid. Contour threading algorithms can be used to prepare output for a pen plotter. Examples of trend surfaces computed for the untransformed zinc data are given in Figure 5.4.

The advantage of trend surface analysis is that it is a technique that is superficially easy to understand, at least with respect to the way the surfaces are calculated. Broad features of the data can be modelled by low-order trend surfaces, but it becomes increasingly difficult to ascribe a physical meaning to complex, higher polynomials. The surfaces are highly susceptible to edge effects, waving the edges to fit the points in the centre of the area, with the result that second order and higher surfaces may reach ridiculously large or small values just outside the area covered by the data. Because it is a general interpolator, the trend surfaces are very susceptible to outliers in the data. Trend surfaces are smoothing functions, rarely passing exactly through the original data points unless these are few and the order of the surface is large. It is implicit in multiple regression that the residuals from a regression line or surface are normally distributed independent errors. The deviations from a trend surface are almost always to some degree spatially dependent; in fact one of the most fruitful uses of trend surface analysis has been to reveal parts of a study area that show the greatest deviation from a general trend (Burrough *et al.* 1977; Davis 1986). The main use of trend surface analysis then, is not as an interpolator within a region, but as a way of removing broad features of the data prior to using some other local interpolator.

The significance of a trend surface The statistical significance of a trend surface can be tested by using the technique of analysis of variance to partition the variance between the trend and the residuals from the trend. Let n be the number of observations, so there are $(n - 1)$ degrees of variation associated with the total variation. The degrees of freedom for regression, m, are determined by the number of terms in the polynomial regression equation, excluding the b_0 coefficient.

For a linear regression, $z(x,y) = b_0 + b_1x + b_2y$, the degrees of freedom for regression $m = 2$. The degrees of freedom for the deviations from the regression are given by $(n - 1) - m$. Table 5.4 presents the terms in the analysis of variance; the reader will note that this table is very similar to Table 5.1 for classification.

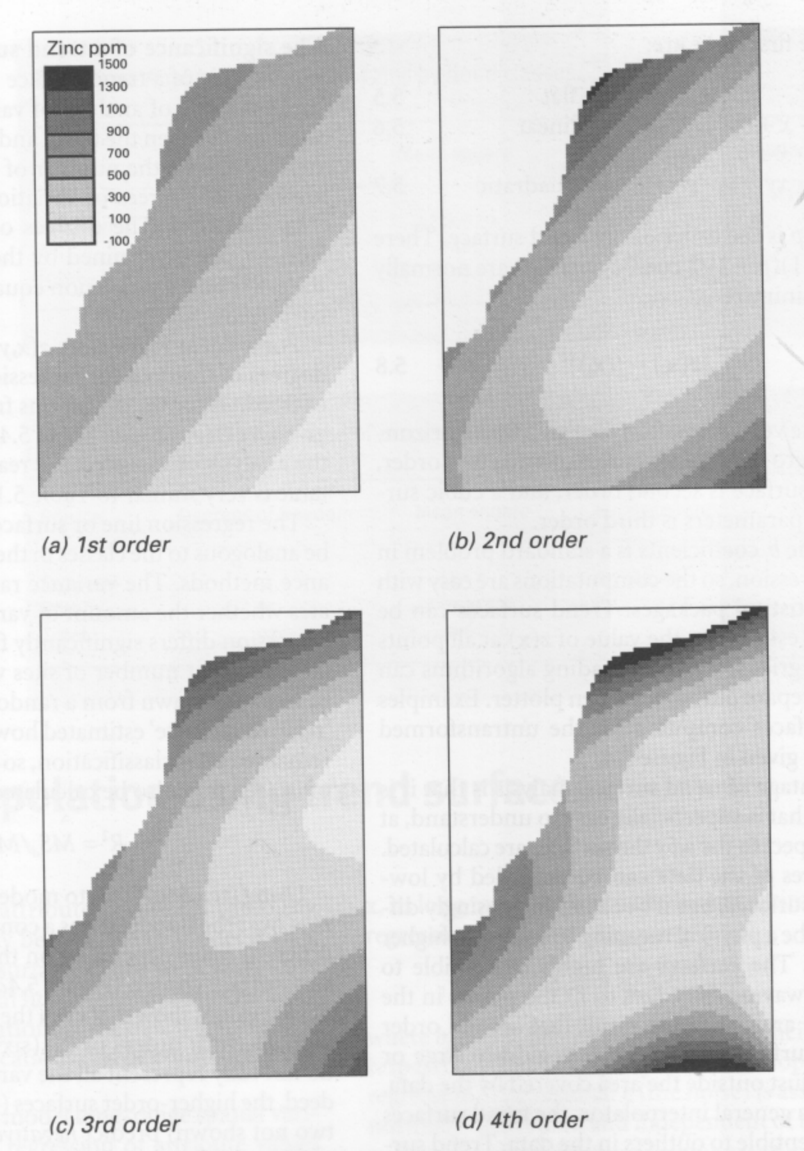
The regression line or surface can be considered to be analogous to the classes in the usual analysis of variance methods. The variance ratio, or 'F-test', estimates whether the amount of variance taken up by the regression differs significantly from that expected for an equivalent number of sites with the same degrees of freedom drawn from a random population. Just as 'relative variance' estimated how much of the variance remained after classification, so a regression goodness of fit (R^2) may also be calculated as:

$$R^2 = MS_d/MS_t \qquad 5.9$$

Using trend surfaces to model the concentration of zinc over the floodplain as a continuous surface yields different maps depending on the order of the regression surface chosen (Figure 5.4*a–d*). The goodness of fit (R^2) values show that even the higher-order surfaces with 21 (fifth order) or 28 (sixth order) coefficients do not fully represent all the variation in the data. Indeed, the higher-order surfaces (4th, 5th, 6th—the last two not shown) predict *negative* values of zinc in the south-eastern corner of the area, which is clearly a serious distortion of reality.

Order:	1	2	3	4	5	6
R^2	.183	.475	.560	.687	.767	.802

Although the R^2 values improve with the order of the regression, at first sight we have no way of judging whether increasing the order of the polynomial significantly increases the fit of the data. The significance of improvement can also be estimated using an analysis of variance, as shown in Box 5.3. Even if significantly better fits can be obtained with higher-order polynomials it is not physically sensible to choose a trend surface that has no physical explanation.



(a) 1st order (b) 2nd order

(c) 3rd order (d) 4th order

Figure 5.4. Simple global trend surfaces for untransformed zinc data

Table 5.4. Analysis of variance terms for linear regression.

Source	Sums of squares	Degrees of freedom	Mean square	Variance ratio
Regression	SS_r	m	MS_r	MS_r/MS_d
Deviation	SS_d	$n-m-1$	MS_d	
Total	SS_t	$n-1$	MS_t	

Box 5.3. ESTIMATING THE SIGNIFICANCE OF USING A HIGHER DEGREE POLYNOMIAL IN TREND SURFACES

Significance testing for increasing the degree of the polynomial

The general ANOVA table for comparing the significance of improved fits is:

source	*sums of squares*	*degrees of freedom*	*mean square*	*variance ratio*
Regression degree p + 1	SS_{rp+1}	m	MS_{rp+1}	MS_{rp+1}/MS_{dp+1}
Deviation from p + 1	SS_{dp+1}	$n-m-1$	MS_{dp+1}	
Regression degree p	SS_{rp}	k	MS_{rp}	MS_{rp}/MS_{dp}
Deviation from p	SS_{dp+1}	$n-k-1$	MS_{dp}	
Improved regression due to increase in order from p to p + 1	SS_{ri} $SS_{rp+r}SS_{rp}$	$m-k$	MS_{ri}	MS_{ri}/MS_{dp+1}
Total	SS_t	$n-1$	MS_t	

As an example, to test for significant improvement in fit from order 2 to order 3 we compute:

ORDER 2 source	sums of squares	degrees of freedom	mean square	variance ratio
Regression	7320729	5	1464145	16.62
Deviation	8106885	92	88118	
Total	15427614	97		

ORDER 3 source	sums of squares	degrees of freedom	mean square	variance ratio
Regression	8639688	9	959965	12.45
Deviation	6787926	88	77135	
Total	15427614	97		

Both surfaces are significant according to a F-test on the variance ratio. To see if there is a significant improvement when increasing the order of the regression surface from two to three we compute:

source	sums of squares	degrees of freedom	mean square	variance ratio
Regression deg p + 1	8639688	9	959965	12.45
Deviation from p + 1	6787926	88	77135	
Regression deg p	7320729	5	1464145	16.62
Deviation from p	8106885	92	88118	
Improved regression due to increase in order	1318959	4	329739	4.27

The variance ratio of 4.27 (degrees of freedom 4,88) is larger than the 1% tabulated value of 3.56 so we conclude that the third order surface is significantly better.

Spatial prediction using global regression on cheap-to-measure attributes

Consideration of Figures 5.2 and 5.4 shows that for the example data set there is a clear geographical relation between the zinc content of the soil and the distance to the river. From other studies (e.g. Leenaers *et al.* 1989*a*, 1989*b*) it is known that heavy metal pollutants in floodplain soils depend on several factors, the most important of which is the distance to the source (i.e. the river) and the elevation of the floodplain. In this area, coarse polluted sediments are deposited with sand in the river levees and fine polluted sediments settle out in low-lying areas where flooding is frequent and inundation persists for longer periods. Areas less frequently flooded receive a smaller pollutant load, as the example of mapping using flood frequency classes showed. Because the attributes *Distance to the river* and *floodplain elevation* are cheap to map, it is possible that we could improve the spatial predictions of zinc content if we could derive an empirical model of the relation between zinc and these independent variables.

The regression model is of the form

$$z(x) = b_0 + b_1P_1 + b_2P_2 + \varepsilon \qquad 5.10$$

where $b_0 \dots b_n$ are regression coefficients and $P_1 \dots P_n$ are independent properties. In this case P_1 is *Distance to river* and P_2 is *Elevation.*

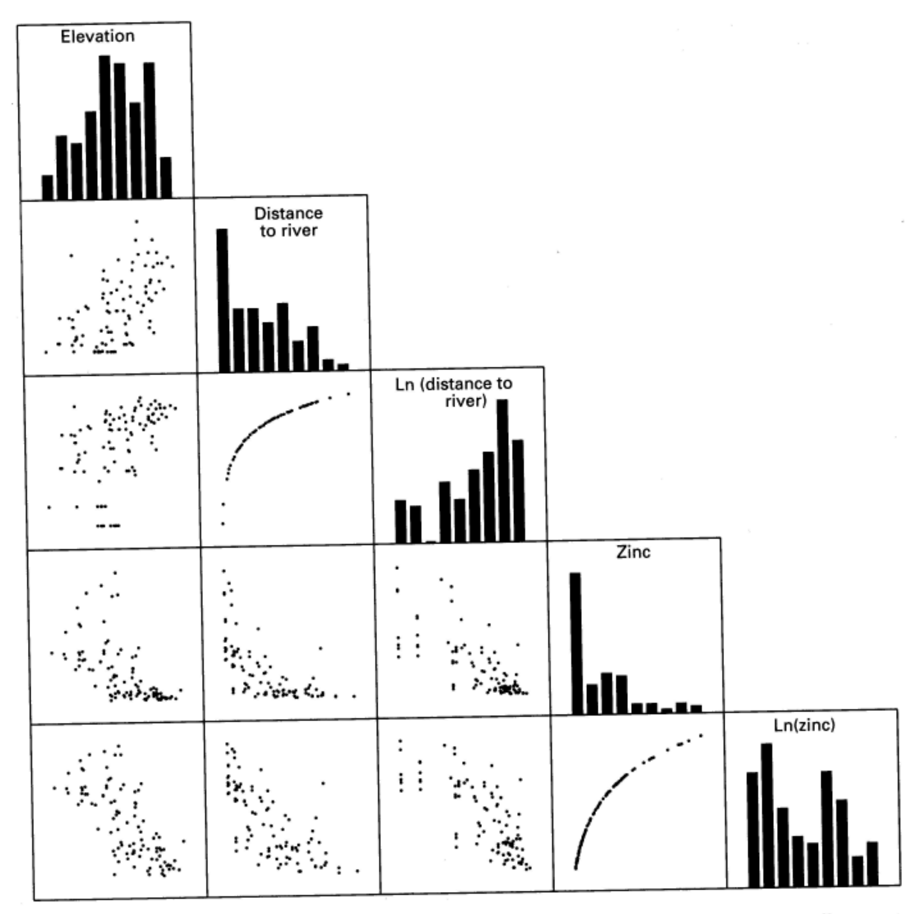


Figure 5.5. Scattergrams of independent and dependent terms in regression of zinc on distance to river and floodplain elevation

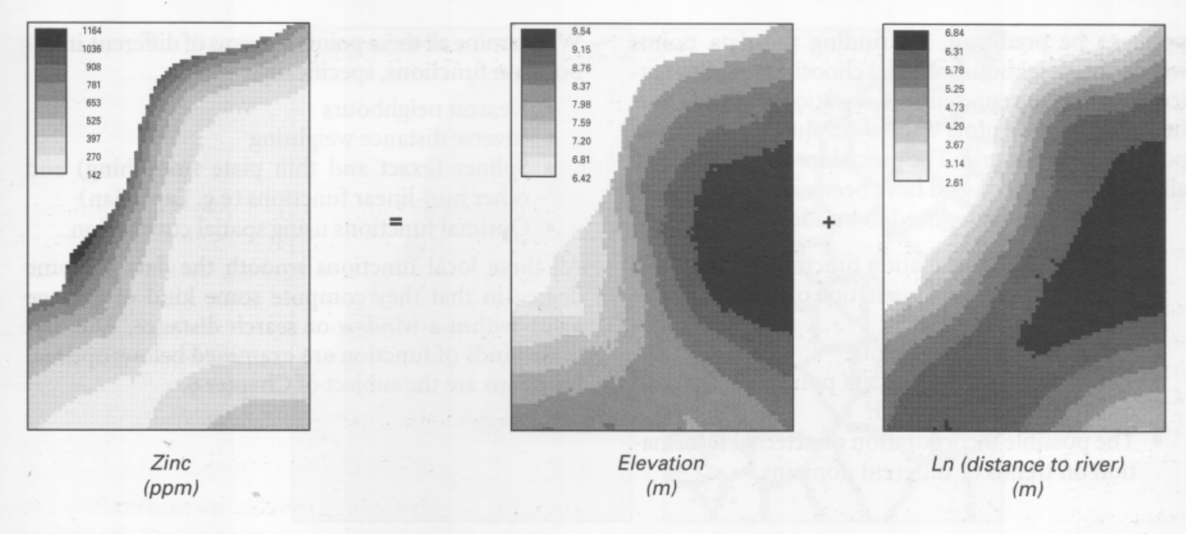


Figure 5.6. Results of computing zinc levels by regression from maps of floodplain elevation and distance to river

Figure 5.5 presents histograms for the independent variables of *Distance to the river* (DRiver—metres), *floodplain elevation* (elevation—metres), and the dependent zinc concentration (ppm); the histograms of logarithmically DRiv and zinc are also included. Inspection of Figure 5.5 indicates that it is best to use the logarithmically transformed variables in the multiple regression. Methods of multiple regression are available in most standard statistical packages.

For the 98 data points of the Meuse zinc example, multiple regression of ln(zinc) against elevation and ln(Distance to river) gives the following relation:

$$\begin{aligned} \ln(\text{zinc}) = {} & 10.000 - 0.292(\text{elevation}) \\ & - 0.333(\text{lnD.river}) \\ & + 0.394(\text{resid error}) \qquad (5.11) \\ R^2 = {} & .749 \end{aligned}$$

Such a regression model is often called a *transfer function* (Bouma and Bregt 1989); it can be computed easily in most GIS (see Chapter 7). The source maps can be obtained in several ways and may be in vector or raster format (see Chapters 7 and 8). Figure 5.6 shows the results.

The same procedure may be used with many other sets of independent and dependent variables, such as temperature and altitude, rain with distance from the sea, vegetation composition as a function of moisture surplus, number of clients and income levels, etc. Geographical coordinates and associated attributes can be combined in one regression to use as much information from the data as possible. The most important point to consider is that the regression model makes physical sense. Note that all regression transfer models are inexact interpolators.

Local, deterministic methods for interpolation

All the methods presented so far have imposed an external, global spatial structure on the interpolation. In all cases short-range, local variations have been dismissed as random, unstructured noise. Intuitively, this is not sensible as one expects the value at an unvisited point to be similar to values measured close by. Therefore people have sought local methods of interpolation that use the information from the nearest data points directly. For this approach, interpolation involves (*a*) defining a search area or neighbourhood around the

point to be predicted, (*b*) finding the data points within this neighbourhood, (*c*) choosing a mathematical function to represent the variation over this limited number of points and (*d*) evaluating it for the point on a regular grid. The procedure is repeated until all the points on the grid have been computed.

The following issues need to be addressed:

- The kind of interpolation function to use.
- The size, shape, and orientation of the neighbourhood.
- The number of data points.
- The distribution of the data points: regular grid or irregularly distributed?
- The possible incorporation of external information on trends or different domains.

We examine all these points in terms of different interpolation functions, specifically:

- Nearest neighbours
- Inverse distance weighting
- Splines (exact and thin plate smoothing) and other non-linear functions (e.g. Laplacian)
- Optimal functions using spatial covariation.

All these local functions smooth the data to some degree in that they compute some kind of average value within a *window* or search distance. The first three kinds of function are examined below; optimal functions are the subject of Chapter 6.

Nearest neighbours: Thiessen (Dirichlet/Voronoi) polygons

Thiessen (otherwise known as Dirichlet or Voronoi) polygons take the classification model of spatial prediction to the extreme whereby the predictions of attributes at unsampled locations are provided by the nearest single data point. Thiessen polygons divide a region up in a way that is totally determined by the configuration of the data points, with one observation per cell. If the data lie on a regular square grid, then the Thiessen polygons are all equal, regular cells with sides equal to the grid spacing; if the data are irregularly spaced, then an irregular lattice of polygons results (Figure 5.7). The lines joining the data points show the Delaunay triangulation, which is the same topology as a TIN (Chapter 3).

Thiessen polygons are often used in GIS and geographical analysis as a quick method for relating point data to space; a common, but implicit use of Thiessen polygons is the assumption that the meteorological data for any given site can be taken from the nearest climate station. Unless there are many observations (which usually there are not), this assumption is not really appropriate for gradually varying phenomena like rainfall and temperature and air pressure because (*a*) the form of the final map is determined by the distribution of the observations, and (*b*) the method maintains the choropleth map fiction of homogeneity within borders and all change at borders. As there is only one observation per tile, there is no way to estimate within-tile variability, short of taking replicate observations.

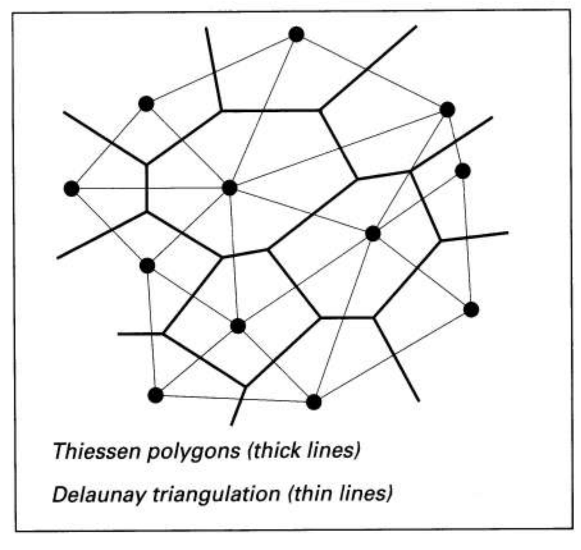


Figure 5.7. An example of a Thiessen polygon net and the equivalent Delaunay triangulation

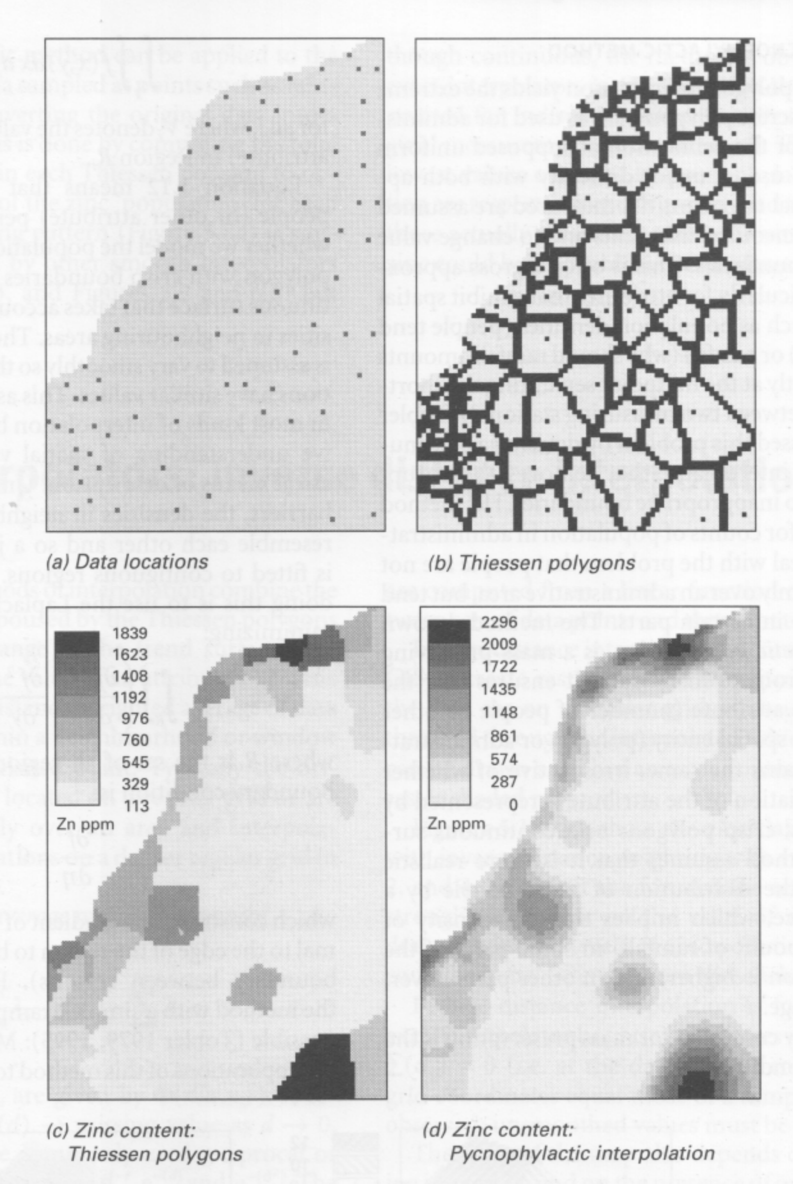


Figure 5.8. (*a*) data locations, (*b*) Thiessen polygons, (*c*) zinc levels per Thiessen polygon for the zinc data, (*d*) pycnophylactic interpolation of zinc from the Thiessen polygon data

An advantage of Thiessen polygons is that they can be easily used with qualitative data like vegetation classes or land use if all you need is a choropleth map and do not mind the strange geometrical pattern of the boundaries. Because all predictions equal the values at the data points, Thiessen polygons are exact predictors.

Figure 5.8*a–c* demonstrates the use of Thiessen polygons for mapping the zinc content of the floodplain soil. Instead of computing averages for externally defined units, the spatial variation of zinc over the floodplain is mapped simply by assigning the measured values to their closest neighbouring cells (Figure 5.8*b,c*). The Thiessen polygon map suggests that the zinc map obtained from information on flooding frequency polygons masks considerable spatial variation of zinc levels, particularly within the flood frequency classes 2 and 3 (cf. Figure 5.2*c*).

Tobler's pycnophylactic method

The Thiessen polygon interpolation yields the extreme case of the discrete polygon map as used for administrative units or the delineation of supposed uniform areas of land use. A major difficulty with both approaches is that the quantities measured are assumed to be homogeneous within units and to change value only at the boundaries. This is often a gross approximation, particularly for attributes that exhibit spatial contiguity, such as population densities (people tend to congregate) or rainfall (why should rainfall amounts change abruptly at the midpoint separating the shortest distance between two measuring stations?). Tobler (1979) addressed this problem by devising a continuous, smooth interpolator that removes the abrupt changes due to inappropriate boundaries. His method was designed for counts of population in administrative areas to deal with the problem that people are not spread uniformly over an administrative area, but tend to congregate in certain parts. The method, known as *pycnophylactic interpolation*, is a mass-preserving reallocation from primary data. It ensures that the *volume* of the attribute (number of people or other attribute) in a spatial entity (polygon or administrative area) remains the same, irrespective of whether the global variation of the attribute is represented by homogeneous, crisp polygons or a continuous surface. The method assumes that it is more realistic to represent the distribution of those people by a smooth surface, which implies that the density of people (or amount of rainfall) in some parts of the area in question is higher, and in other parts lower, than the average.

The primary condition for mass preservation is the invertibility condition:

$$\int_{R_i}\int f(x,y)\,dx\,dy = V_i \qquad 5.12$$

for all i, where V_i denotes the value (population, count, attribute) in Region R_i.

Equation 5.12 means that the total volume of people (or other attribute) per polygon is invariate whether we model the population count by a uniform polygon with crisp boundaries or by a smooth, continuous surface that takes account of population densities in neighbouring areas. The constraining surface is assumed to vary smoothly so that neighbouring locations have similar values. This assumption is common in most kinds of interpolation because it links intuitive understanding of spatial variation to mathematical means of description. Unless there are physical barriers, the densities in neighbouring areas tend to resemble each other and so a joint, smooth surface is fitted to contiguous regions. The simplest way of doing this is to use the Laplacian condition, i.e. by minimizing:

$$\int_{R}\int \left(\frac{\partial f^2}{\partial x} + \frac{\partial f^2}{\partial y}\right) dx\,dy \qquad 5.13$$

where R is the set of all regions. The most general boundary condition is:

$$\frac{\partial f}{\partial \eta} = 0 \qquad 5.14$$

which constrains the gradient of the fitted surface normal to the edge of the region to be flat (η indicates the boundary between regions). Figure 5.9 illustrates the method with a simple example. Other options are possible (Tobler 1979, 1995): Martin (1996) reviews the applications of this method to the 1991 UK census.

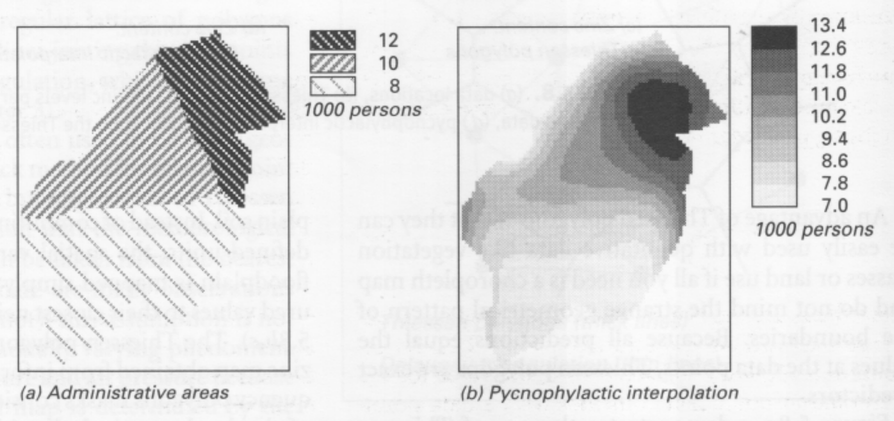


Figure 5.9. Population counts displayed by (*a*) uniform for administrative areas and (*b*) by pycnophylactic interpolation

The pycnophylactic method can be applied to the zinc data or other data sampled at points such as rainfall amounts, by converting the original data into a density function. This is done by computing the total amount of zinc within each Thiessen polygon to obtain a pseudo count of the zinc 'population' for each polygon. The resulting pattern (Figure 5.8*d*) is similar to that obtained by other smooth interpolators (see sections 5.2, 5.3, and Figures 5.9 and 6.6) but though continuous, the method is obviously not an exact interpolator, in the sense that the values interpolated for individual grid cells differ from the original measurements at the same cell. The highest and lowest values obtained by pycnophylactic interpolation are respectively much greater, and much less, than those actually measured, which for sampled data like zinc could bring problems of interpretation.

Linear interpolators: inverse distance interpolation

Inverse distance methods of interpolation combine the ideas of proximity espoused by the Thiessen polygons with the gradual change of the trend surface. The assumption is that the value of an attribute z at some unvisited point is a distance-weighted average of data points occurring within a neighbourhood or window surrounding the unvisited point. Typically the original data points are located on a regular grid or are distributed irregularly over an area and interpolations are made to locations on a denser regular grid in order to make a map.

Weighted moving average methods compute

$$\hat{z}(\mathbf{x}_0) = \sum_{i=1}^{n} \lambda_i \cdot z(\mathbf{x}_i) \quad \sum_{i=1}^{n} \lambda_i = 1 \qquad 5.15$$

where the weights λ_i are given by $\Phi(d(\mathbf{x},\mathbf{x}_i))$. A requirement is that $\Phi(d) \rightarrow$ missing value as $d \rightarrow 0$, which is given by the commonly used reciprocal or negative exponential functions d^{-r}, $e^{-(d)}$ and $e^{-(d^2)}$. The most common form of $\Phi(d)$ is the inverse distance weighting predictor whose form is:

$$\hat{z}(\mathbf{x}_0) = \sum_{i=1}^{n} z(\mathbf{x}_i) \cdot d_{ij}^{-r} \Big/ \sum_{i=1}^{n} d_{ij}^{-r} \qquad 5.16$$

where the x_j are the points where the surface is to be interpolated and the x_i are the data points. Because in equation 5.16, $\Phi(d) \rightarrow \infty$ as $d \rightarrow 0$, the value for an interpolation point that coincides with a data point must be simply copied over. The simplest form of this is called the *linear interpolator*, in which the weights are computed from a linear function of distance between sets of data points and the point to be predicted.

Inverse distance interpolation is commonly used in GIS to create raster overlays from point data. Once the data are on a regular grid, contour lines can be threaded through the interpolated values and the map can be drawn as either a vector contour map or as a raster shaded map.

Figures 5.10*a*,*b*,*c* show the zinc data interpolated with inverse distance weighting with values of r or 1, 2, and 4. Note that the form of the resulting map is strongly dependent on the value of r. Conclusions drawn about the levels of zinc pollution when $r = 1$ could be quite different from those when $r = 4$.

Inverse distance interpolation is forced to be an exact interpolator because it produces infinities when $\sum(d_{ij}) = 0$ (i.e. at the data points) so if the output grid coordinates equal those of a sampling point the observed, unsmoothed values must be copied over.

The form of the map also depends on the clustering in the data and on the presence of outliers. Inverse distance interpolations commonly have a 'duck-egg' pattern around solitary data points with values that differ greatly from their surroundings, though this can be modified to a certain extent by altering the search criteria for the data points to account for anisotropy. The method has no inbuilt method of testing for the quality of predictions so the map quality can only be assessed by taking extra observations. Note that these must be for the same support as the original observations, though it is arguable that since inverse distance interpolation smooths within a zone proportional to the value of r, this is the true resolution of the interpolation and not the sample support.

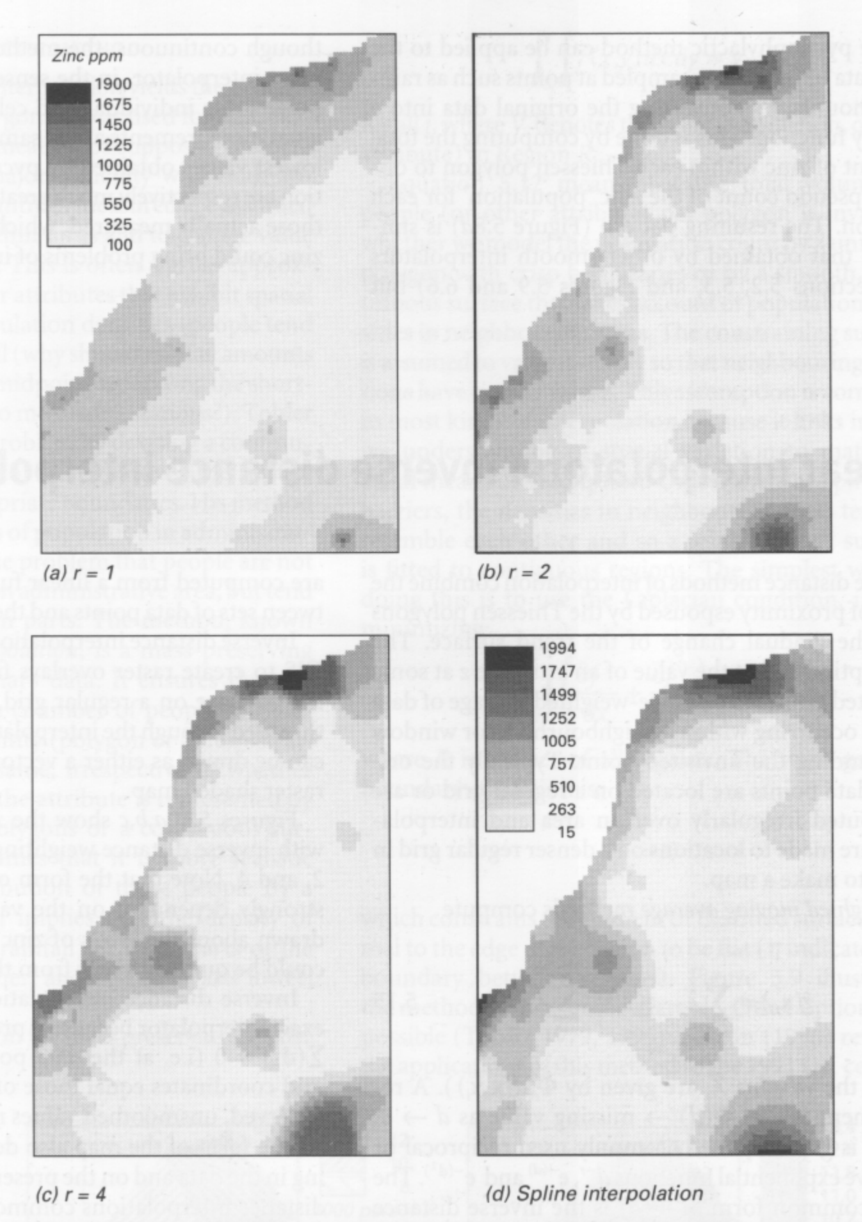


Figure 5.10. (*a–c*) Inverse distance interpolation showing the effect of the weighting parameter on the results; (*d*) thin plate spline mapping of zinc

Splines

Before computers were used to fit curves to sets of data points, draughtsmen used flexible rulers to achieve the best locally fitting smooth curves by eye. The flexible rulers were called splines. The flexible rulers were held in place by weights on pegs at the data points while the line was drawn (Pavlidis 1982). The modern equivalent is the plastic coated flexible ruler sold in most office equipment shops. It can be shown that the line

drawn along a spline ruler is approximately a piece-wise cubic polynomial that is continuous and has continuous first and second derivatives.

Spline functions are mathematical equivalents of the flexible ruler. They are piece-wise functions, which is to say that they are fitted to a small number of data points exactly, while at the same time ensuring that the joins between one part of the curve and another are continuous. This means that with splines it is possible to modify one part of the curve without having to recompute the whole, which is not possible with trend surfaces (Figure 5.10).

The general definition of a piece-wise polynomial function $p(x)$ is:

$$p(x) = p_i(x) \quad x_i < x < x_{i+1} \quad i = 0, 1, \ldots, k-1 \qquad 5.17$$

$$p^j(x_i) = p^j_{i+1}(x_i) \quad j = 0, 1, \ldots, r-1; \quad i = 1, 2, \ldots, k-1 \qquad 5.18$$

The points $x_i, \ldots. x_{k-1}$ that divide an interval x_0, x_k into k sub-intervals are called *break points* and the points of the curve at these values of x are commonly called *knots* (Pavlidis 1982). The functions $p_i(x)$ are polynomials of degree m or less. The term r is used to denote the constraints on the spline. When $r = 0$, there are no constraints on the function; when $r = 1$ the function is continuous without any constraints on its derivatives. If $r = m + 1$, the interval x_0, x_k can be represented by a single polynomial, so $r = m$ is the maximum number of constraints that leads to a piece-wise solution. For $m = 1$, 2, or 3, a spline is called linear, quadratic, or cubic. The derivatives are of order 1, 2, $m - 1$, so a quadratic spline must have one continuous derivative at each knot, and a cubic spline must have two continuous derivatives at each knot. For a simple spline where $r = m$ there are only $k + m$ degrees of freedom. The case of $r = m = 3$ has particular significance because the term *spline* was first used for cubic piece-wise polynomial functions. The term *bicubic spline* is used for the three-dimensional situation where surfaces instead of lines need to be interpolated. Equation 5.18 means that adjacent polynomials have the same value at a break point for a given $j = r$.

Because of difficulties of calculating simple splines over a wide range of separate sub-intervals, such as might be the case with a digitized line, most practical applications use a special kind of spline called B-splines. B-splines are themselves the sums of other splines that by definition have the value of zero outside the interval of interest (Pavlidis 1982). B-splines, therefore, allow local fitting from low-order polynomials in a simple way.

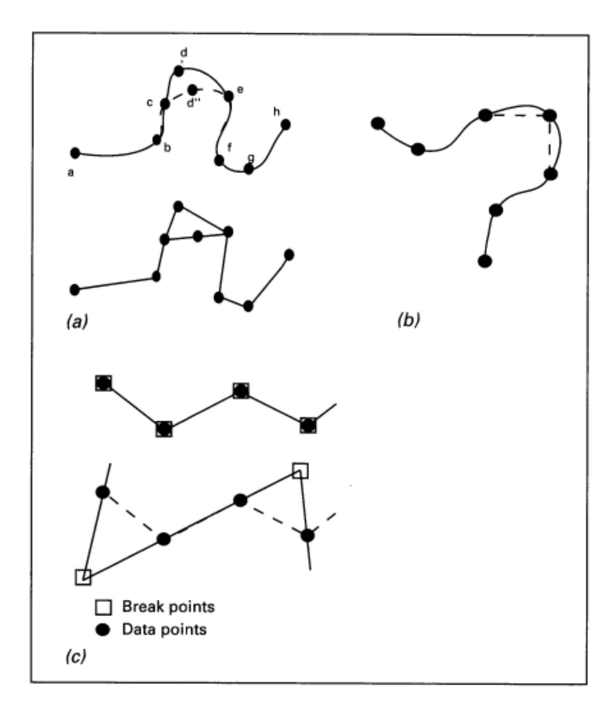


Figure 5.11. Some properties of splines (*a*) local adjustments mean local changes; (*b*) exact splines round off sharp corners; (*c*) choosing to locate break points at or between data points has a large effect on the resulting spline

B-splines are often used for smoothing digitized lines for display, such as the boundaries on soil and geological maps where cartographic conventions expect smooth, flowing lines. Pavlidis notes that complex shapes such as those occurring in text fonts, can be more economically defined in terms of B-splines than in sets of data points. The use of B-splines for smoothing polygon boundaries, however, can lead to certain complications, particularly when computing areas and perimeters. If the area of a polygon is calculated from digitized data points using the trapezoidal rule (Box 3.3), it will be different from the area that results from smoothing the boundaries with B-splines. Another problem may arise when high-order B-splines are used to smooth sinuous boundaries that also include sharp, rectangular corners (Figure 5.11*b*).

A problem with using splines for interpolation is whether one should choose the break points to coincide with the data points or to be interleaved with them. Different results for the interpolated spline may result from the two approaches (Figure 5.11*c*). Note that with splines the maxima and minima do not necessarily occur at the data points.

Thin plate splines as surface interpolators

Exact splines are commonly used in GIS and drawing packages for smoothing curves and contour lines to improve appearance where it is assumed that exact interpolation is required. When data have been sampled in two or three dimensions, effects of natural variation and measurement errors may be such that an exact spline may produce local artefacts of excessively high or low values. These artefacts can be removed by using *thin plate splines*, in which the exact spline surface is replaced by a locally smoothed average. As with line interpolation, spline methods of surface interpolation assume that an approximate function should pass as close as possible to the data points while also being as smooth as possible.

When the data contain a source of random error, i.e.:

$$y(x_i) = z(x_i) + \varepsilon(x_i) \qquad 5.19$$

where z is the measured value of an attribute at point x_i, and epsilon is the associated random error. The spline function $p(x)$ should pass 'not too far' from the data values and so the smoothing spline is the function f that minimizes

$$A(f) + \sum_{i=1}^{n} w_i^2[f(x_i) - y(x_i)]^2 \qquad 5.20$$

The term $A(f)$ represents the 'smoothness' of the function f, and the second term represents its 'proximity' or 'fidelity' to the data. The weights w_i^2 are chosen to be inversely proportional to the error variance;

$$w_i^2 = p/\mathrm{Var}[\varepsilon(x_i)] = p/s_i^2 \qquad 5.21$$

where the value of p reflects the relative importance given by the user to each characteristic of the smoothing splines.

Smoothing, thin plate splines are frequently used for interpolating elevation to create digital elevation models where it is necessary to interpolate large areas quickly and efficiently—see Hutchinson 1995, Mitasova *et al.* 1995. Both texts demonstrate their use for multivariate interpolation of attributes such as annual mean rainfall from a trivariate spline function of longitude, latitude, and elevation.

Using splines for the zinc data Figure 5.10*d* shows the variation of zinc as mapped by splines. Note that the spline surface tends to 'draw up' the areas around sample sites with large zinc values.

Advantages and disadvantages of splines Because splines are piece-wise functions using few points at a time, the interpolating values can be quickly calculated. Test data for smooth surfaces show predictions are very close to the values being interpolated, providing the measurement errors associated with the data are small (Mitasova *et al.* 1995). In contrast to trend surfaces and weighted averages, splines retain small-scale features. Both linear and surficial splines are aesthetically pleasing and quickly produce a clear overview of the data. The smoothness of splines means that mathematical derivatives can easily be calculated for direct analysis of surface geometry and topology (see Chapter 8 and Figure 5.14*a*). The incorporation of linear parametric sub-models (regression models) makes the interpolation of dependent variables on point supports easy.

Some disadvantages of splines have already been mentioned. Others are that there are no direct estimates of the errors associated with spline interpolation, though these may be obtained by a recursive technique known as 'jack-knifing' (Dubrule 1984). The most critical disadvantage may be that thin plate splines provide a view of reality that is unrealistically smooth; in some situations, such as the estimation of attribute values for numerical models, this property could generate misleading results.

A comparison of simple global and local methods

We have explained several different kinds of interpolation technique, but how do they compare? Which produces the most consistent results, when should one method be preferred to another? To give a first insight, Table 5.5 compares all methods in terms of the minimum and maximum interpolated values, and the proportions of the area estimated to be above the thresholds of 500, 1000, and 1500 ppm zinc.

Table 5.5. Summary of results of deterministic interpolation

Method	Minimum value (ppm)	Maximum value (ppm)	Per cent area >500 ppm	Per cent area >1000 ppm	Per cent area >1500 ppm
Flood frequency (untransformed)	206	770	11.96	0.00	0.00
Trend surface—order 1	−98	791	38.53	0.00	0.00
Trend surface—order 2	109	1350	31.88	3.37	0.00
Trend surface—order 3	−13	1256	33.26	3.66	0.00
Trend surface—order 4	−100	1500	28.30	5.88	0.69
Regression on distance + elevation	142	1164	17.82	0.64	0.00
Thiessen polygons	113	1839	28.11	9.65	5.31
Pycnophylactic method	0	2296	28.35	10.15	3.87
Inverse distance $r = 1$	136	1541	28.92	1.16	0.03
Inverse distance $r = 2$	114	1827	28.82	3.99	0.61
Inverse distance $r = 4$	113	1839	28.47	7.05	2.49
Thin plate splines	15	1994	30.44	6.74	1.87

Comparing these results we see that global trend surfaces of orders higher than 3 are unreliable. Trend surfaces are therefore really only useful for describing broad geographical trends (cf. Burrough *et al.* 1977). Pycnophylactic methods and thin plate splines stretch the maximum and minimum values above and below the recorded values but inverse distance with a $r = 2$ gives results closest to the original data. The range of predictions of the areas exceeding the three arbitrary limits produced by all these methods is greatest for the values exceeding 1000 ppm; without independent data and a review of the performance of the methods on other data sets it is impossible to say which method is generally the best. The inverse distance and thin plate splines seem to produce the most 'natural' looking surfaces, but this may just be a reflection of our cultural preferences, as these surfaces are always much smoother than the underlying reality.

Digital elevation models as a special case of continuous surfaces created by interpolation

This section examines the digital elevation model (DEM) as a special case of an interpolated, continuous surface that has many uses in GIS. DEMs were originally computed as a precursor of the orthophoto map, but today they have many other applications (Box 5.4), including the eye-catching ability to display spatial data over the underlying landform (e.g. Plate 1.4). This section explains how DEMs are created and how they can be modelled in the computer, leading to derived products such as block diagrams and digital orthophotos. Other products, created from the mathematical derivatives of DEMs are explained in Chapter 8 as part of the spatial analysis of continuous surfaces.

Methods of representing DEMs

The variation of surface elevation over an area can be modelled in many ways. DEMs can be represented

Box 5.4. USES OF DEMS

Some common uses of digital elevation models

- Storage of elevation data for digital topographic maps in national databases
- Creation of digital and analogue orthophoto maps
- Cut and fill problems in road design and other civil and military engineering projects
- Three-dimensional display of landforms for military purposes (weapon guidance systems, pilot training) and for landscape design and planning (landscape architecture).
- For analysis of cross-country visibility (also for military and for landscape planning purposes).
- For planning routes of roads, locations of dams, etc.
- For statistical analysis and comparison of different kinds of terrain.
- Source data for derived maps of maps, aspect, profile curvature, shaded relief insolation and hydrological and ecological modelling – see Chapter 8.
- As a background for displaying thematic information or for combining relief data with thematic data such as soils, land use, or vegetation.
- Provide data for simulation models of landscapes and landscape processes.

either by mathematically defined surfaces or by point or line images. Line data can be used to represent contours and profiles, and critical features such as streams, ridges, shorelines, and breaks in slope. In GIS, DEMs are modelled by regular grids (*altitude matrices*) and triangular irregular networks (*TINs*). The two forms are inter-convertible and the preference for one or the other depends on the kind of data analysis that needs to be carried out.

Altitude matrices are the most common form of discretized elevation surface. Originally they were derived from quantitative measurements of stereoscopic aerial photographs made on analytical stereoplotters such as the GESTALT GPM-II (Kelly *et al.* 1977). The DEM was a by-product needed to produce scale-correct *Orthophoto maps* (see below). Alternatively, the altitude matrix can be produced by interpolation from irregularly or regularly spaced data points in the same way as other quantitative data.

Because of the ease with which matrices can be handled in the computer, in particular in raster-based geographical information systems, the altitude matrix has become the most available form of DEM. Britain, Australia, the United States of America and indeed much of the world, are completely covered by coarse matrices derived from 1 : 250 000 scale topographic maps (grid cell sides of 1–5 km cell size are available on the Internet). Higher-resolution matrices based on 1 : 50 000 or 1 : 25 000 maps and aerial photography are becoming increasingly available for these and other countries.

Figures 5.12 and 5.13 give examples of altitude matrix DEMs. Figure 5.12 shows a DEM with a cell size of 30 m × 30 m. Elevation data may be displayed in several ways—Figure 5.12 shows a grey scale image of simple elevation and a grey scale image of surface aspect (Chapter 8 explains how this is derived). By choosing an appropriate grey scale, the image appears to be illuminated from the north (top), thereby providing a sense of three-dimensional relief. Figure 5.13 shows a similar image from another area, but now the aspect information is 'draped' over the DEM to yield a *block diagram* that portrays the relief naturally. Altitude matrices are the starting-point for deriving much useful information about landform, such as slope, profile convexity, solar irradiance, lines of sight, and surface topology, as explained in Chapter 8.

THE TRIANGULAR IRREGULAR NETWORK (TIN)

Although altitude matrices are useful for calculating contours, slope angles and aspects, hill shading and automatic basin delineation (see Chapter 8), the regular grid system is not without its disadvantages. These disadvantages include (*a*) the large amount of data redundancy in areas of uniform terrain (*b*) the inability to adapt to areas of differing relief complexity without changing the grid size, (*c*) the exaggerated

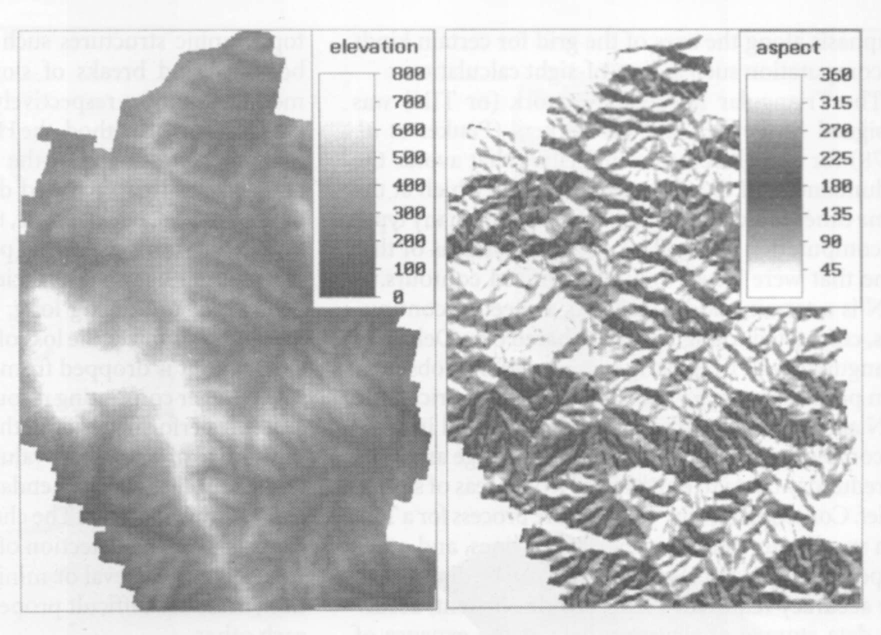


Figure 5.12. (Left) Grey-scale altitude matrix (pixel size 30 × 30 m); (right) Aspect map shaded to enhance effect of shaded relief. Data courtesy A. Skidmore

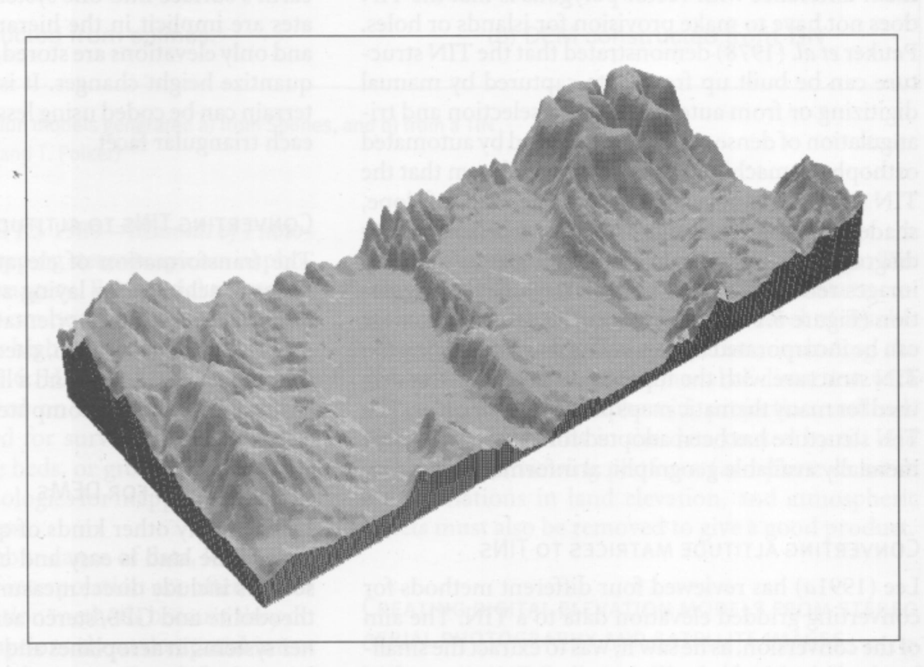


Figure 5.13. Shaded relief data draped over a block diagram creates a strong impression of the relief (data courtesy S. M. de Jong)

emphasis along the axes of the grid for certain kinds of computation such as line-of-sight calculations.

The Triangular Irregular Network (or TIN) was designed by Peuker and co-workers (Peuker *et al.* 1978) for digital elevation modelling that avoids the redundancies of the altitude matrix and which at the same time would also be more efficient for many types of computation (such as slope) than systems of that time that were based only on digitized contours. A TIN is a terrain model that uses a sheet of continuous, connected triangular facets based on a Delaunay triangulation of irregularly spaced nodes or observation points (Figure 5.7). Unlike altitude matrices, the TIN allows extra information to be gathered in areas of complex relief without the need for huge amounts of redundant data to be gathered from areas of simple relief. Consequently, the data capture process for a TIN can specifically follow ridges, stream lines, and other important topological features that can be digitized to the accuracy required. TINs provide efficient, accurate data storage of elevation data at the expense of introducing a triangular discretization that may hinder some kinds of spatial analysis, such as the derivation of surface geometry and topology.

TINs are modelled with a topological vector structure similar to those used for polygon networks, and the TIN data structure was explained in Chapter 3. The main difference with vector polygons is that the TIN does not have to make provision for islands or holes. Peuker *et al.* (1978) demonstrated that the TIN structure can be built up from data captured by manual digitizing or from automated point selection and triangulation of dense raster data gathered by automated orthophoto machines. They have also shown that the TIN structure can be used to generate maps of slope, shaded relief, contour maps, profiles, horizons, block diagrams, and line of sight maps, though the final map images retain an imprint of the Delaunay triangulation (Figure 5.14*b*). Information about surface cover can be incorporated by overlaying and intersecting the TIN structure with the topological polygon structure used for many thematic maps of discrete variables. The TIN structure has been adopted for at least one commercially available geographical information system.

Converting altitude matrices to TINs

Lee (1991*a*) has reviewed four different methods for converting gridded elevation data to a TIN. The aim of the conversion, as he saw it, was to extract the smallest possible set of irregularly spaced elevation points that provides a maximum of information about topographic structures such as peaks, ridges, valley bottoms, and breaks of slope. He evaluated three methods (known respectively as the VIP—'Very important points' method, the HT—hierarchy transform methods, and the DH—the drop heuristic method) and showed that each had different advantages and disadvantages. For example, the VIP method was best at locating critical points (peaks or pits), while the HT method has a very efficient data structure at the expense of producing long, thin triangles. The DH method minimizes the loss of information every time a data point is dropped from the grid, but it requires much larger computing resources. Lee found that the absolute performance of all three methods depends on the choice of parameter values and so was unable to make absolute recommendations as to which algorithm should be used. The choice seems to depend on the user's aims—detection of extremes, efficient data storage, and retrieval or minimizing loss of information; these are difficult properties to trade off against each other.

Recently Dutton (1996) has proposed an interesting alternative compact method for modelling planetary relief based on recursive polyhedral tesselation of a sphere into equilateral triangular facets. The method is known as the Geodetic Elevation Model of Planetary Relief because it attempts to bring the whole of the earth's surface into one system. Horizontal coordinates are implicit in the hierarchy of nested triangles and only elevations are stored, using single bit flags to quantize height changes. It is claimed that an entire terrain can be coded using less than one bit of data for each triangular facet.

Converting TINs to altitude matrices

The transformation of elevation data from TIN to raster is achieved by laying a regular grid of the required resolution and orientation over the TIN. Each grid cell is visited in turn, checked to see in which are the nearest TIN apices and a linear or bicubic average of the TIN heights is computed.

Data sources for DEMs

Unlike many other kinds of quantitative data, elevation of the land is easy and cheap to measure. Data sources include direct measurement in the field with theodolite and GPS, stereo aerial photographs, scanner systems in aeroplanes and satellites, and the digitizing of contour lines on paper maps. For systematic mapping, elevation data are derived by the methods

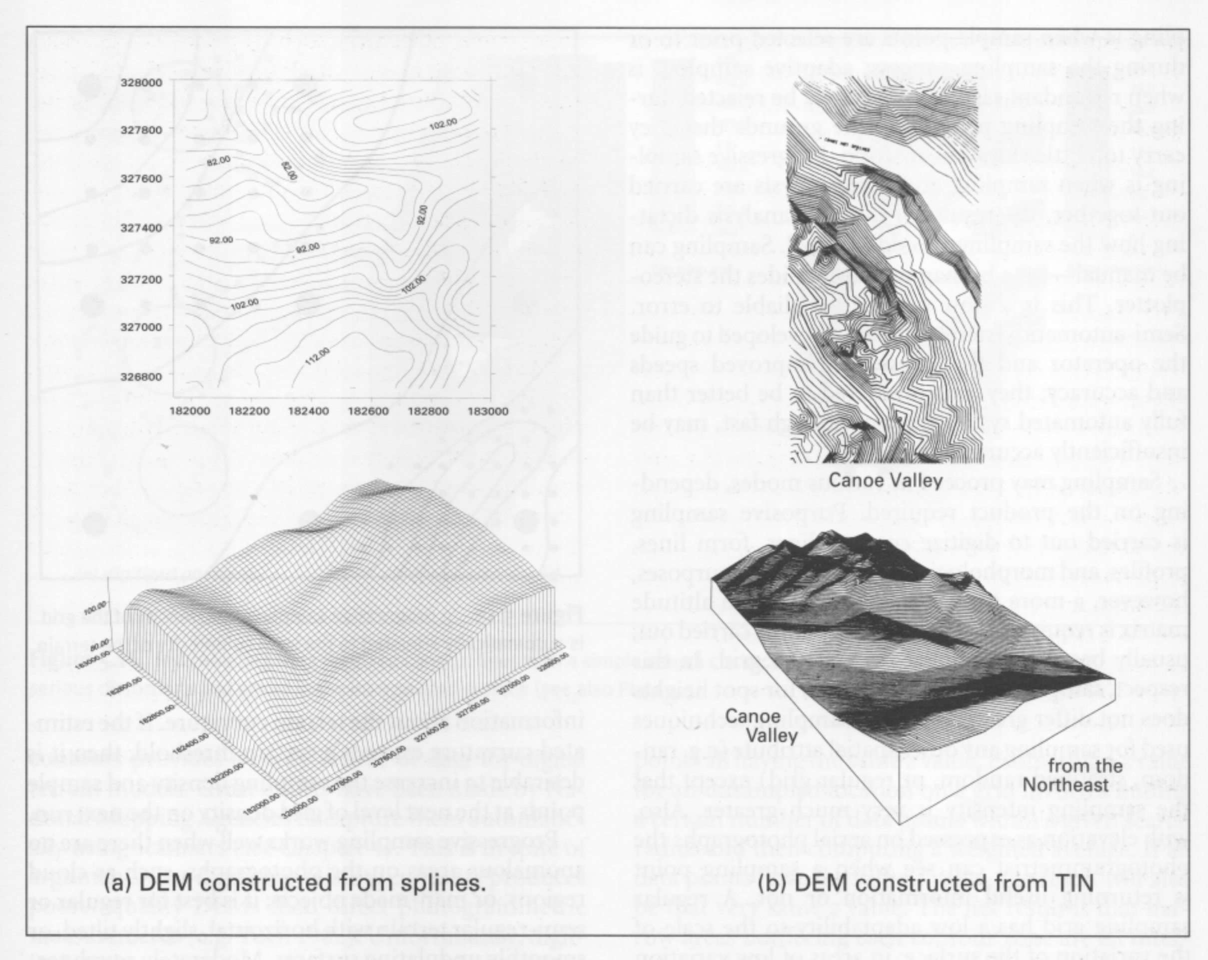


Figure 5.14. Digital elevation models generated a) from Splines, and b) from a TIN (data courtesy of A. de Roo and T. Poiker)

of *photogrammetry* (ASPRS 1980—*Manual of Photogrammetry*) from overlapping stereoscopic aerial photographs and satellite imagery. For special purposes, other kinds of scanners can be used, such as airborne laser interferometry for high-accuracy surface measurements (Plates 3.5, 3.6; Hazelhoff, pers. comm.). Sonar scanners mounted on boats, submarines, or hovercraft are also used for surveying the elevation patterns of sea and lake beds, or ground-penetrating radar and seismic technologies for mapping the elevation of sub-surface layers.

As there is often an abundance of data, local, simple (linear) methods of interpolation are often better than complex interpolation methods because they do not need to make assumptions about the spatial interactions and they are quick to compute. When stereo aerial photographs and satellite images are the source of elevation data we have complete coverage of the landscape at the level of resolution of the image. The creation of a DEM is then the extraction of data to create a proper *hypsometric surface* at a level of resolution appropriate for the application, including the geometric correction and removal of distortion with respect to the chosen spheroid, projection, and orientation. Distortions in the data caused by tilt and wobble in the viewing platform (aeroplane or satellite), variations in land elevation, and atmospheric effects must also be removed to give a good product.

Creating Digital Elevation Models from Stereo Aerial Photography and Satellite Images

Makarovic (1976) distinguished several methods of photogrammetric sampling for DEMs. Selective sam-

pling is when sample points are selected prior to or during the sampling process; adaptive sampling is when redundant sample points may be rejected during the sampling process on the grounds that they carry too little extra information. Progressive sampling is when sampling and data analysis are carried out together, the results of the data analysis dictating how the sampling should proceed. Sampling can be manual—i.e. a human operator guides the stereo-plotter. This is a slow process and liable to error. Semi-automatic systems have been developed to guide the operator and these result in improved speeds and accuracy; they are considered to be better than fully automated systems, which though fast, may be insufficiently accurate.

Sampling may proceed in various modes, depending on the product required. Purposive sampling is carried out to digitize contour lines, form lines, profiles, and morphological lines. For many purposes, however, a more general DEM based on an altitude matrix is required, and so areal sampling is carried out, usually based on a regular or irregular grid. In this respect, sampling aerial photographs for spot heights does not differ greatly from the sampling techniques used for sampling any other spatial attribute (e.g. random, stratified random, or regular grid) except that the sampling intensity is very much greater. Also, with elevation as expressed on aerial photographs the photogrammetrist can see when a sampling point is returning useful information or not. A regular sampling grid has a low adaptability to the scale of the variation of the surface; in areas of low variation too many points may be sampled, and in areas of large variation the number of sample points may be too few. If the operator is given the freedom to make observations at will, the sampling can be highly subjective. Makarovic (1973) proposed a method called 'progressive sampling' that provides an objective and automatable method for sampling terrain of varying complexity in order to produce an altitude matrix.

Progressive sampling involves a series of successive runs, beginning first with a coarse grid, and then proceeding to grids of higher densities (Figure 5.15). The grid density is doubled on each successive sampling run, and the points to be sampled are determined by a computer analysis of the data obtained on the preceding run.

The computer analysis proceeds as follows: a square patch of nine points on the coarsest grid is selected and the height differences between each adjacent pair of points along the rows and columns is calculated. The second differences are then calculated. These carry information about the terrain curvature. If the estimated curvature exceeds a certain threshold, then it is desirable to increase the sampling density and sample points at the next level of grid density on the next run.

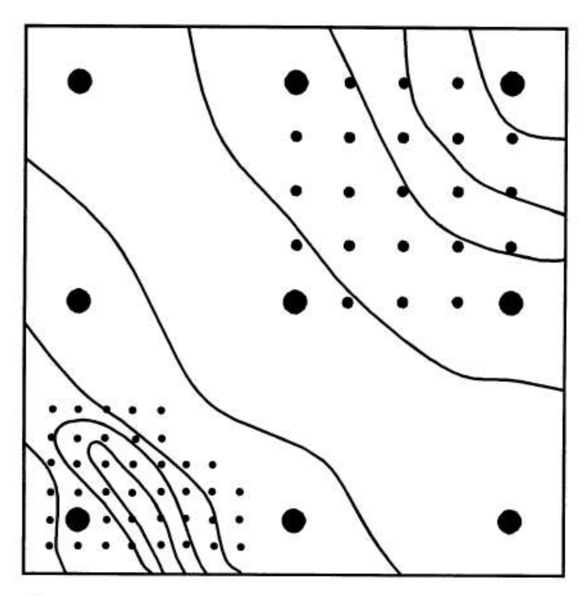


Figure 5.15. In progressive sampling, the density of the grid is automatically adjusted to the local complexity of the terrain

Progressive sampling works well when there are no anomalous areas on the photographs, such as cloud regions, or man-made objects; it is best for regular or semi-regular terrain with horizontal, slightly tilted, or smoothly undulating surfaces. Moderately rough terrain with distinct morphological features and some anomalous areas can be better handled by a modification of progressive sampling called composite sampling (Makarovic 1977). In composite sampling, abrupt steps in the terrain or the boundaries of natural or anomalous objects are first delineated by hand before sampling within these areas. Rough terrain types with many abrupt changes may not be efficiently covered by any semi-automated progressive or composite sampling approach, and all data may have to be gathered by selective sampling.

Finally, the data collected by progressive and composite sampling must be automatically converted to fill the whole altitude matrix uniformly.

Interpolating from Digitized Contours to an Altitude Matrix

The most common line model of terrain is the set of contour lines on printed maps, and digitizing the

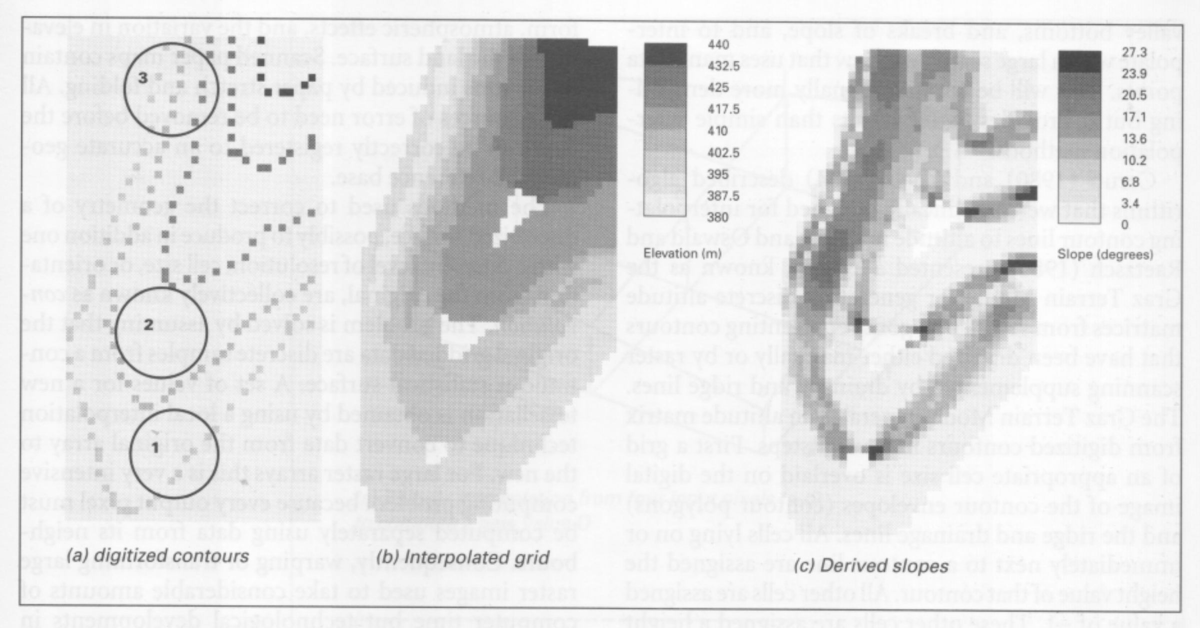


Figure 5.16. Interpolating from digitized contour lines with a simple search circle can create serious distortions and errors in the interpolated surface (see also Plate 4.5)

contours provides a ready source of data for digital terrain models. Great efforts have been made by National Mapping Agencies to capture them automatically using scanners (see Chapter 4). This is in spite of arguments that digitizing existing contours produces poorer quality DEMs than direct photogrammetric measurements (e.g. Yoeli 1982). Unfortunately, digitized contours are not especially suitable for computing slopes or for making shaded relief models and so they must be converted to an altitude matrix.

Unsatisfactory results are often obtained when people attempt to create their own DEMs by digitizing contours and then using local interpolation methods like inverse distance weighting and kriging (see next chapter) to interpolate the digitized contours to a regular grid. The problem is not so much the assumptions behind the computation of the interpolation weights as in the geometry of the search algorithms Apart from the problem of assuming that smooth contour lines are true representations of terrain and locational errors are only caused by paper stretch and digitizer placement, interpolating digitized contours to a regular grid can result in the creation of severe artefacts in the resulting surface. Curiously enough, the more care that is taken to digitize a contour line with many sampled points, the greater the problem.

The reasons for the errors can be seen in Figure 5.16. Accurately digitized contour lines return sets of data points all having the same *z* value. Estimating a *z* value for an unsampled location on a grid involves finding a certain number of data points within a given search radius and then computing a weighted average. If all data points have the same *z* value, the result will also be that very same *z* value. The net result is that narrow areas bordering each contour zone are all interpolated with the same *z* value so that each contour is transformed into a 'padi' (or rice) terrace. The problem is usually greater in areas of low relief where contour lines are further apart and the chance that the search algorithm only catches data from one contour line is greatest.

Of course, other errors result from the stepped 'padi' DEMs; computing slopes (Chapter 8) often yields a map with unnatural tiger stripes (Plate 4.5). Very often these errors are overlooked, because their origin is not understood. For example, the Walker data set used by Isaaks and Srivastava (1989) appears to be corrupted by errors resulting from interpolating from digitized contours as evidenced by banding in their figures 5.8 and 5.9*c* and 5.10*c*.

A better solution to contour line problems is to use another interpolation method that is designed to deal with the kinds of data yielded by digitizing contours. If this is not available, the best strategy is to thin the digitized contour points to the very minimum, to add extra points to the data set to indicate peaks, ridges,

valley bottoms, and breaks of slope, and to interpolate with a large search window that uses many data points. This will be computationally more demanding but it provides better results than simple interpolation methods.

Ceruti (1980) and Yoeli (1984) described algorithms that were specifically designed for interpolating contour lines to altitude matrices and Oswald and Raetzsch (1984) presented a system, known as the Graz Terrain Model for generating discrete altitude matrices from sets of polygons representing contours that have been digitized either manually or by raster scanning supplemented by drainage and ridge lines. The Graz Terrain Model generates an altitude matrix from digitized contours in several steps. First a grid of an appropriate cell size is overlaid on the digital image of the contour envelopes (contour polygons) and the ridge and drainage lines. All cells lying on or immediately next to a contour line are assigned the height value of that contour. All other cells are assigned a value of −1. These other cells are assigned a height value in the following step which is a linear interpolation procedure working within rectangular subsets or windows of the raster database. Interpolation usually takes place along four search lines oriented N–S, E–W, NE–SW, NW–SE. The interpolation proceeds by computing the local steepest slope for the window as a simple function of the difference between the heights of cells that already have a height value assigned. For each window the slopes are grouped in four classes. Beginning with the class of steepest slope, unassigned cells within the window are assigned a height; the procedure is repeated for the other slope classes excepting flat areas which are computed separately after all steep parts of the DEM have been computed. Oswald and Raetzsch (1984) claim that this interpolation method, the 'sequential steepest slope algorithm', is a robust and useful technique. Note that commercial GIS may include procedures for interpolating from contours but do not disclose the algorithms, so the user has no idea of the quality of the result. Carrara *et al.* (1997) report a comparison of several widely-available methods for generating DEM's from digitized contours in which commercial methods performed well.

Geometry correction for altitude matrices and other raster data

All scanned imagery collected from airborne and space platforms contains geometrical distortions because of the curvature of the earth, tilt and wobble in the platform, atmospheric effects, and the variation in elevation of the land surface. Scanned paper maps contain distortions induced by paper stretch and folding. All these sources of error need to be removed before the data can be correctly registered to an accurate geometrical reference base.

The methods used to correct the geometry of a discretized surface, possibly to produce in addition one with a different level of resolution, cell size, or orientation from the original, are collectively known as *convolution*. The problem is solved by assuming that the original gridded data are discrete samples from a continuous statistical surface. A set of values for a new tessellation is obtained by using a local interpolation technique to convert data from the original array to the new. For large raster arrays this is a very intensive computing problem because every output pixel must be computed separately using data from its neighbours. Consequently, warping or transforming large raster images used to take considerable amounts of computer time but technological developments in computer processors and memories have reduced the time required to rotate a large raster array from several hours to a fraction of a second. Warping involves two separate processes. The first is the computation of the addresses of the output pixels relative to the input data matrix. The general equation for a first order warp (translation, rotation, scale) given by Adams *et al.* (1984) is:

$$u = a_0 + a_1 a_2 x + a_1 a_3 y$$
$$v = b_0 + b_1 b_2 x + b_1 b_3 y \qquad 5.22$$

where x,y are the original pixel coordinates, u,v are the new coordinates, a_0 and b_0 are translation values, a_1 and b_1 are x and y scale values respectively, and a_2, b_2 and a_3, b_3 are dependent on the angle of rotation θ as given by

$$a_2 = \cos\theta \quad b_2 = -\sin\theta$$
$$a_3 = \sin\theta \quad b_3 = \cos\theta \qquad 5.23$$

For warps that are not coplanar, such as in the problem of fitting satellite imagery to the curved surface of the earth to match a conventional map projection such as the Universal Transverse Mercator, higher-order warps must be used. Several methods are used for the local interpolation. The simplest, and most limited, is to interpolate the cell value on the warped surface from its closest neighbour on the original surface. A better alternative is a bilinear interpolater in which the new value is computed from the four input pixels surrounding the output pixel (Figure 5.17*a*). The best interpolator is probably the cubic convolution

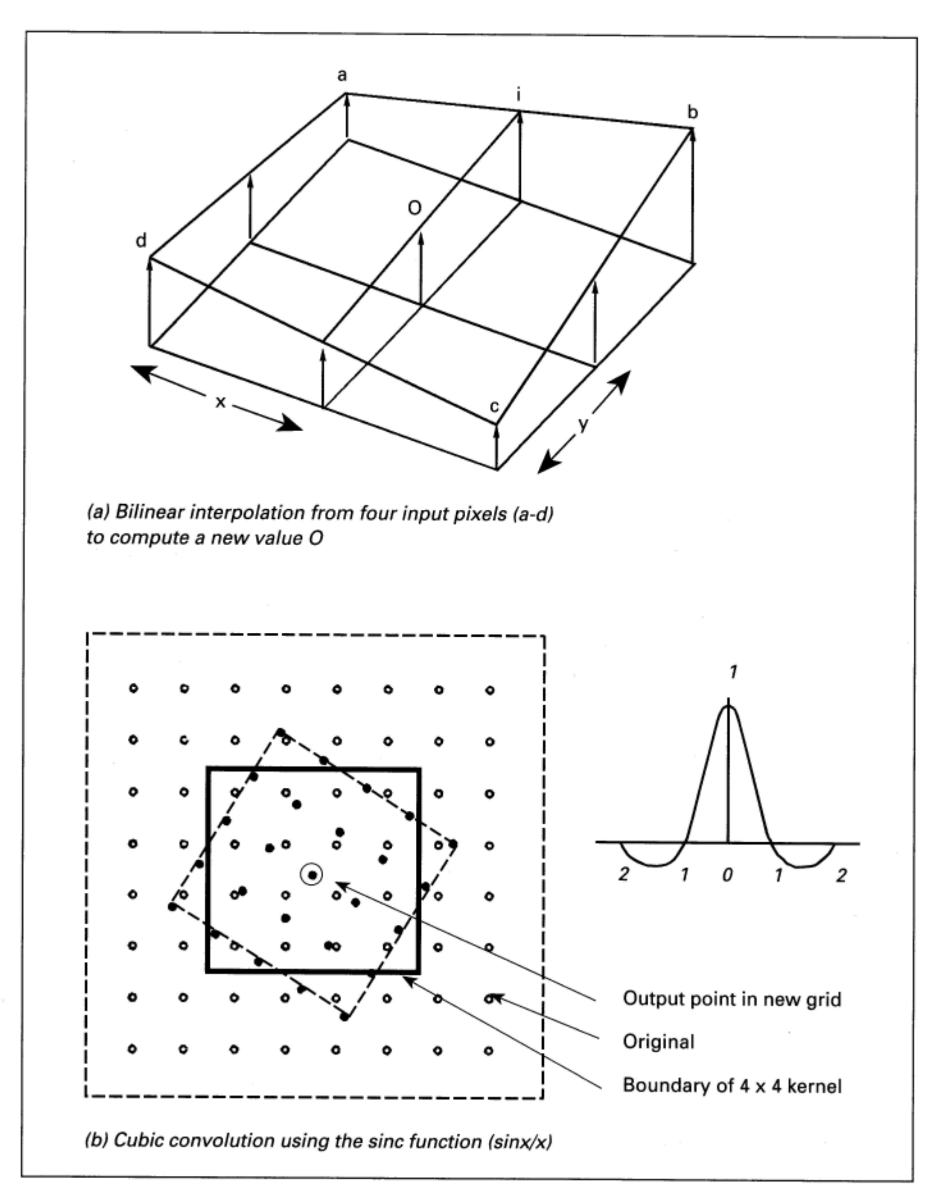


Figure 5.17. (*a*) Bilinear interpolation from four input pixels to compute a new value at cell **O**; (*b*) cubic convolution using the sinc function (sin*x* / *x*)

(Figure 5.17*b*) which uses a neighbourhood of 16 pixels and weighted sum approach based on a two-dimensional version of the sinc $x(\sin x/x)$ function.

DIGITAL ORTHOPHOTOS

Chapter 8 explains how many useful products can be derived from altitude matrices, but perhaps the most useful digital cartographic product that can be obtained from a DEM is the modern digital orthophoto. Digital orthophoto maps are being used increasingly to provide geometrically correct, highly detailed photographic images for deriving data on land use and land cover and as an informative background to data on utilities, municipal administration, and environmental studies (Plates 1.1–1.4).

The Orthophoto is a photo map that is geometrically correct in the same way that a topographical map

is geometrically correct with respect to map scale and projection. Orthophoto maps are photogrammetric products that were designed initially to reduce the costs and to speed up large-scale topographic mapping in areas where full ground surveys had not been carried out or where there were financial constraints for producing full topographic coverage at large map scales or for updating maps at short notice. Unlike topographic maps, in which the terrain features are depicted in standard codes and symbols on a scale-correct base map using a standardized map projection and map legend, orthophotos are aerial photo mosaics that have been geometrically corrected to a standard scale and projection. Orthophoto map scales are usually larger than 1 : 25 000. Orthophoto maps and digital products may carry a limited amount of topographic information such as contour lines, administrative boundaries, and thematic data, which are overprinted on the image. Until recently, orthophotos were only available as paper photomaps, but now they can be provided as very high-resolution raster digital files in 32-bit colour (see Plates 1.1–1.4).

Remotely sensed satellite images (Thematic Mapper, SPOT) can also be corrected for scale and projection and used as surrogates for topographic maps. Because of the coarser spatial resolution the resulting maps are usually at smaller scales than orthophoto maps. Satellite images are currently used in GIS and remote sensing systems to provide background information for thematic data. Analysis of the spectral information in the satellite images can also supply useful and important information about land use, vegetation, soil conditions, hydrology, and so on. Image analysis techniques are used to help identify and classify geographical objects from the spectral information in the images. Very similar analysis techniques can also be used with spectral data obtained from airborne sensors of non-visible radiation (e.g. radar, infra-red) which may give better spatial resolution than the satellite images. Remote sensing is often used as an affordable means of obtaining information about temporal changes on the earth's surface. Remotely sensed images and their classified products are usually primarily available as digital databases; they can also be printed on paper or film using high-quality film writers.

The main difference between the orthophoto map (or satellite image map) and the topographic map is that the former contains photographic information on all aspects of the landscape that are visible at the level of resolution used while the topographic map presents structured, classified information about selected aspects of the landscape. The orthophoto map is easier to use for orientation and object recognition; the topographic map is a rapid key to the structure, location, and connectivity of predefined objects.

Digital orthophoto maps are produced by scanning black and white or colour aerial photographs at a resolution varying between 200μm–10μm. The images are corrected for distortion by mathematical correction methods using information from ground control points (at least six per aerial photo), camera optics, colour balance, and terrain elevation differences, which requires an accurate DEM. The photomosaic is made by merging the corrected digital images from adjacent photographs, adjusting them for differences in grey-scale intensity or in colour balance. Topographic information on administrative boundaries and other features can be added from the GIS database or digitized by hand, and the digital orthophoto can be distributed on CD-ROM or as a hardcopy made with a laser film writer.

Full colour digital orthophoto images are stored with a data accuracy of 32 bits per pixel. The volume of data can be reduced enormously by recoding the image using only 8 bits per pixel and image compression techniques which means that orthophoto images can be easily viewed on personal computers or distributed on the Internet, and can be incorporated as a data layer in most geographical information systems.

Digital Orthophotomaps can be produced at a wide range of scales or perhaps better, levels of resolution. As a rule of thumb, the scale of the map is usually reckoned to be about a maximum of three times the scale of the aerial photographs. Scales of 1 : 1000–1 : 3500 are used for urban areas; scales of 1 : 50 000 can be used for regional mapping.

Updating orthophotos Producing the DEM for the geometric correction of aerial photographs is the most expensive part of making digital orthophotos. However, since in most cases the shape of the landscape does not change over time, the DEM does not need to be recomputed every time new aerial photography is flown, and the existing surface can be used to support the correction of new photography. When the digital data are in a GIS it is very easy to compare the original situation (e.g. land use) with the new situation and quantitative estimates of change can easily be made. In certain situations, such as open-cast mining, the erosion of dunes on the coast, or in civil engineering cut-and-fill operations for construction, DEM updating allows volume changes to be computed quickly

and accurately. If imagery is available for several years the whole process of change can be studied, which could be of value in coastal or fluviatile areas.

Much structured topographic data can be collected directly from aerial photographs, particularly for large to medium-scale applications. Structured data on invisible aspects of the landscape, such as buried pipelines and surveyed parcel boundaries, must be collected by field survey. Modern GIS permit structured vector data to be viewed and analysed with raster data, such as scanned data from satellite images or digital orthophotos. Consequently the structured topographic data, which shows an abstract view of the world, can be seen in its proper context. This can be useful for checking on data quality and correctness and for seeing the relations between structured data and other aspects of the location.

Questions

1. Compare and contrast the different methods for creating spatial classifications in physical sciences such as hydrology, soil science, and geology with those encountered in the human sciences, such as demography, epidemiology, and politics.

2. Explore the assumptions inherent in the ANOVAR method and how these affect the integrity of the classification method of spatial prediction.

3. Explain the assumptions behind trend surface analysis and show how these may seriously affect the quality of the results. Consider residual errors, outliers, assumptions of stationarity (homogeneity of variation) and independence. Does a large value of R^2 necessarily mean that unsampled points will be predicted correctly? How would you check the quality of the predictions independently?

4. Discuss the kinds of spatial process encountered in natural and social sciences (physical and human geography, ecology, hydrology and meteorology, demography, etc.) that lead to spatial variation that can be modelled by a global trend or regression surface.

5. Explain how you would set up a method of objective testing that would compare the predictive quality of different interpolation techniques, such as inverse distance weighting versus thin-plate splines or regression surfaces.

6. Explain why the TIN may be unsuitable for modelling the continuous variation of physical attributes other than elevation measured at point locations.

7. Explore ways of generalizing to three dimensions the interpolation methods presented in this chapter.

Suggestions for further reading

Hutchinson, M. F. (1995). Interpolating mean rainfall using thin plate smoothing splines. *International Journal of Geographical Information Systems*, 9: 385–404.

Mitasova, H., Mitas, L., Brown, W. M., Gerdes, D. P., Kosinovsky, I., and Baker, T. (1995). Modelling spatially and temporally distributed phenomena: new methods and tools for GRASS GIS. *International Journal of Geographical Information Systems*, 9: 433–46.

Tobler, W. (1979). Smooth pycnophylactic interpolation for geographic regions. *J. American Statistical Association*, 74(367): 519–36.

SIX

Optimal Interpolation Using Geostatistics

When data are abundant, most interpolation techniques give similar results. When data are sparse, however, the assumptions made about the underlying variation that has been sampled and the choice of method and its parameters can be critical if one is to avoid misleading results. Geostatistical methods of interpolation, popularly known as *kriging*, attempt to optimize interpolation by dividing spatial variation into three components—(*a*) deterministic variation (different levels or trends) that can be treated as useful, soft information, (*b*) spatially autocorrelated, but physically difficult to explain variations, and finally (*c*) uncorrelated noise. The character of the spatially correlated variation is encapsulated in functions such as the autocovariogram and (semi) variogram, and these provide the information for optimizing interpolation weights and search radii. Experimental variograms are computed from sample data in one, two, or three spatial dimensions. These experimental data are fitted by one of a limited number of variogram models, which serve to provide data for computing interpolation weights.

Geostatistical methods provide great flexibility for interpolation, providing ways to interpolate to areas or volumes larger than the support (block kriging), methods for interpolating binary data (indicator kriging), and methods for incorporating soft information about trends (universal kriging) or stratification (stratified kriging). All these methods of interpolation yield smoothly varying surfaces accompanied by an estimation variance surface. In contrast to smooth interpolators, the methods of conditional simulation, given the variogram and the original observations, provide the best estimators of data values at unsampled points. The resulting surfaces need not be smooth, but have lower estimation errors than kriging predictions. Combining soft information and conditional simulation is useful for computing data for raster-based environmental models.

Finally, the information in the variogram can be used to help optimize sampling schemes for mapping from point data.

A brief introduction to regionalized variable theory and kriging

When comparing all the maps made in Chapter 5 of the zinc concentration in the floodplain soils we see that the different methods return large differences in the global patterns, the amount of local detail, and in the minimum and maximum values predicted. These differences were summed up in Table 5.5 in terms of the estimates given by the various methods of the percentages of the area thought to be above 500, 1000, and 1500 ppm zinc respectively. Clearly, if we were faced with all these different results and must make a decision on the costs of cleaning up soil with more than a certain level of zinc, it would be very difficult to choose. None of the methods of interpolation discussed so far can provide direct estimates of the quality of the predictions made in terms of an estimation variance for the predicted value at unsampled locations. In all cases, the only way to determine the goodness of the predictions would be to compute estimates for a set of extra *validation points* that had not been used in the original interpolation. Apart from research studies to judge the quality of performance of a technique (e.g. Laslett and McBratney 1990*a*, 1990*b*), this is rarely done because it costs money.

A further objection to all methods so far is that there is no *a priori* method of knowing whether the best values have been chosen for the weighting parameters or if the size of the search neighbourhood is appropriate. As we have seen, the control parameters of trend surfaces and inverse distance weighting can be varied to produce quite different maps which in turn provide different estimates of the distribution of the variable being mapped. Moreover, no method studied so far provides sensible information on:

- the number of points needed to compute the local average,
- the size, orientation, and shape of the neighbourhood from which those points are drawn,
- whether there are better ways to estimate the interpolation weights than as a simple function of distance,
- the errors (uncertainties) associated with the interpolated values.

These questions led the French geomathematician Georges Matheron and the South African mining engineer D. G. Krige to develop optimal methods of interpolation for use in the mining industry. The methods are now being increasingly used in groundwater modelling, soil mapping, and related fields, and packages for geostatistical interpolation are becoming important modules of commercial GIS.

Geostatistical methods for interpolation start with the recognition that the spatial variation of any continuous attribute is often too irregular to be modelled by a simple, smooth mathematical function. Instead, the variation can be better described by a stochastic surface. The attribute is then known as a *regionalized variable*; the term applies equally to the variation of atmospheric pressure, elevation above sea level, or the distribution of continuous demographic indicators. Interpolation with geostatistics is known as *kriging*, after D. G. Krige.

Geostatistical methods provide ways to deal with the limitations of deterministic interpolation methods listed above, and ensure that the prediction of attribute values at unvisited points is optimal in terms of the assumptions made. An optimal policy is a rule in dynamic programming for choosing the values of a variable so as to optimize the criterion function (Bullock and Stallybrass 1977). The interpolation methods developed by Matheron, Krige, and their co-workers are optimal in the sense that the interpolation weights are chosen so as to optimize the interpolation function, i.e. to provide a Best Linear Unbiased Estimate (BLUE) of the value of a variable at a given point. The same theory can be used for optimizing sample networks.

Regionalized variable theory assumes that the spatial variation of any variable can be expressed as the sum of three major components (Figure 6.1). These are (*a*) a structural component, having a constant mean or trend; (*b*) a random, but spatially correlated component, known as the variation of the *regionalized variable*, and (*c*) a spatially uncorrelated random noise or residual error term. Let **x** be a position in 1, 2, or 3 dimensions. Then the value of a random variable Z at **x** is given by

$$Z(\mathbf{x}) = m(\mathbf{x}) + \varepsilon'(\mathbf{x}) + \varepsilon'' \qquad 6.1$$

where $m(\mathbf{x})$ is a deterministic function describing the 'structural' component of Z at **x**, $\varepsilon'(\mathbf{x})$ is the term denoting the stochastic, locally varying but spatially dependent residuals from $m(\mathbf{x})$—the *regionalized*

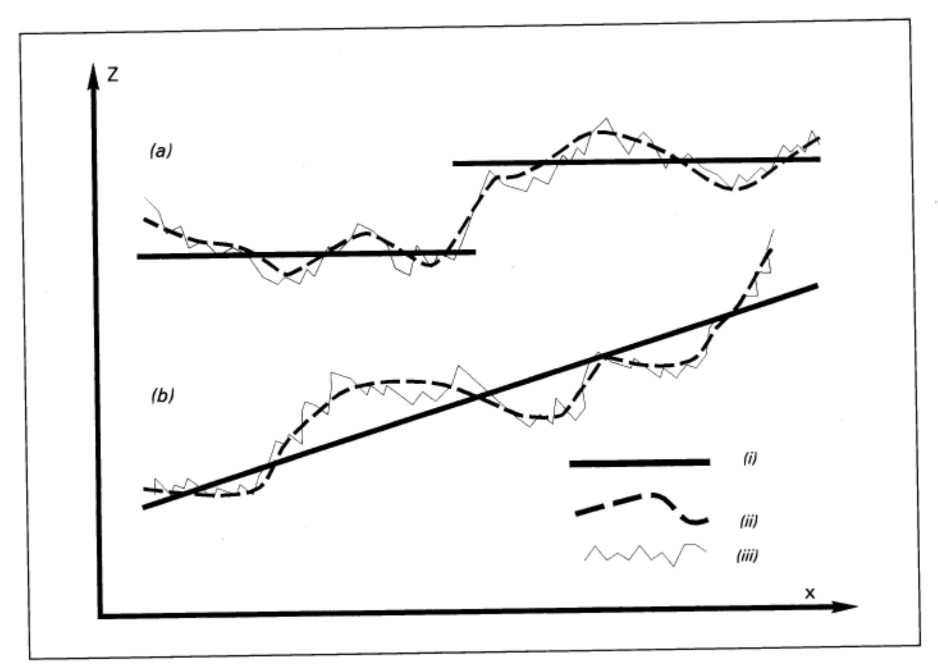


Figure 6.1. Regionalized variable theory divides complex spatial variation into (i) average behaviour such as differences in mean levels (upper) or a trend (lower), (ii) spatially correlated, but irregular ('random') variation, and (iii) random, uncorrelated local variation caused by measurement error and short range spatial variation

variable—, and ε'' is a residual, spatially independent Gaussian noise term having zero mean and variance σ^2. Note the use of the capital letter to indicate that Z is a random function and not a measured attribute z.

The first step is to decide on a suitable function for $m(\mathbf{x})$. In the simplest case, where no trend or *drift* is present, $m(\mathbf{x})$ equals the mean value in the sampling area and the average or expected difference between any two places $\mathbf{x}$ and $\mathbf{x}+\mathbf{h}$ separated by a distance vector $\mathbf{h}$, will be zero:

$$E[Z(\mathbf{x}) - Z(\mathbf{x}+\mathbf{h})] = 0 \qquad 6.2$$

where $Z(\mathbf{x})$, $Z(\mathbf{x}+\mathbf{h})$ are the values of random variable Z at locations $\mathbf{x}$, $\mathbf{x}+\mathbf{h}$. Also, it is assumed that the variance of differences depends only on the distance between sites, $\mathbf{h}$, so that

$$E[\{Z(\mathbf{x}) - Z(\mathbf{x}+\mathbf{h})\}^2] = E[\{\varepsilon'(\mathbf{x}) - \varepsilon'(\mathbf{x}+\mathbf{h})\}^2] = 2\gamma(\mathbf{h}) \qquad 6.3$$

where $\gamma(\mathbf{h})$ is known as the semivariance. The two conditions, stationarity of difference and variance of differences, define the requirements for the *intrinsic hypothesis* of regionalized variable theory. This means that once structural effects have been accounted for, the remaining variation is homogeneous in its variation so that differences between sites are merely a function of the distance between them. We can rewrite equation 6.1 as:

$$Z(\mathbf{x}) = m(\mathbf{x}) + \gamma(\mathbf{h}) + \varepsilon'' \qquad 6.4$$

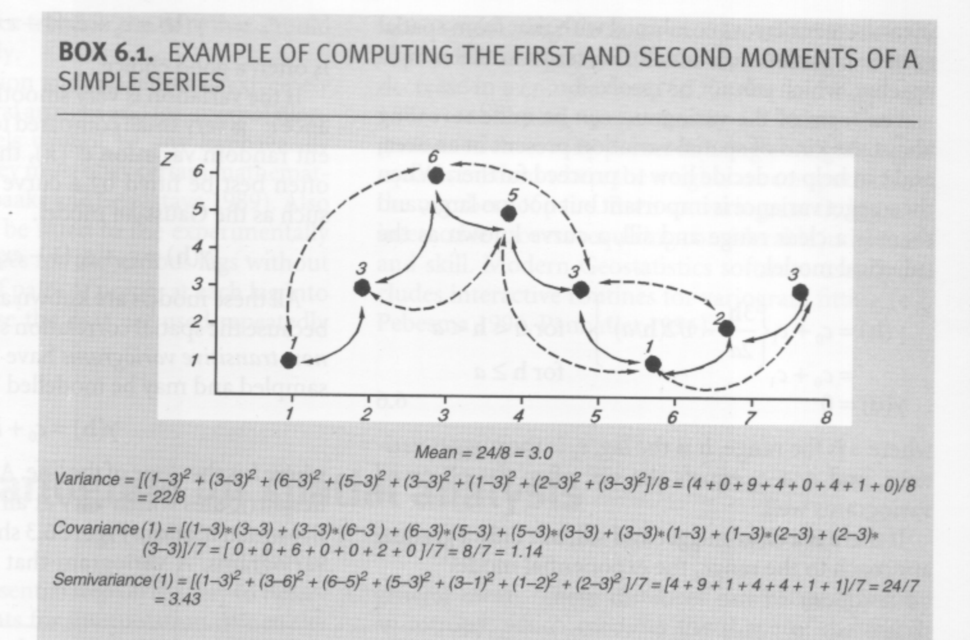
in order to show the equivalence between $\varepsilon'(\mathbf{x})$ and $\gamma(\mathbf{h})$.

If the conditions specified by the intrinsic hypothesis are fulfilled, the semivariance can be estimated from sample data:

$$\hat{\gamma}(\mathbf{h}) = \frac{1}{2n}\sum_{i=1}^{n}\{z(\mathbf{x}_i) - z(\mathbf{x}_i + \mathbf{h})\}^2 \qquad 6.5$$

where n is the number of pairs of sample points of observations of the values of attribute z separated by distance $\mathbf{h}$ (see Box 6.1). A plot of $\hat{\gamma}(\mathbf{h})$ against $\mathbf{h}$ is known as the *experimental variogram.* The experimental variogram is the first step towards a quantitative description of the regionalized variation. The variogram provides useful information for interpolation, optimizing sampling and determining spatial patterns. To do this, however, we must first fit a theoretical model to the experimental variogram.

BOX 6.1. EXAMPLE OF COMPUTING THE FIRST AND SECOND MOMENTS OF A SIMPLE SERIES



$$\text{Mean} = 24/8 = 3.0$$

$$\text{Variance} = [(1-3)^2 + (3-3)^2 + (6-3)^2 + (5-3)^2 + (3-3)^2 + (1-3)^2 + (2-3)^2 + (3-3)^2]/8 = (4+0+9+4+0+4+1+0)/8 = 22/8$$

$$\text{Covariance}(1) = [(1-3)*(3-3) + (3-3)*(6-3) + (6-3)*(5-3) + (5-3)*(3-3) + (3-3)*(1-3) + (1-3)*(2-3) + (2-3)*(3-3)]/7 = [0+0+6+0+0+2+0]/7 = 8/7 = 1.14$$

$$\text{Semivariance}(1) = [(1-3)^2 + (3-6)^2 + (6-5)^2 + (5-3)^2 + (3-1)^2 + (1-2)^2 + (2-3)^2]/7 = [4+9+1+4+4+1+1]/7 = 24/7 = 3.43$$

Fitting variogram models

Figure 6.2 shows a typical experimental variogram of data from a not too smoothly varying attribute, such as a soil property. The curve that has been fitted through the experimentally derived data points displays several important features. First, at large values of the lag, $\mathbf{h}$, it levels off. This horizontal part is known as the *sill*; it implies that at these values of the lag there is no spatial dependence between the data points because all estimates of variances of differences will be invariant with sample separation distance. Second, the curve rises from a low value of $\gamma(\mathbf{h})$ to the sill, reaching it at a value of $\mathbf{h}$ known as the *range*. This is the critically important part of the variogram because it describes how inter-site differences are spatially dependent. Within the range, the closer sites are together the more similar they are likely to be. The range gives us an answer to the question posed in weighted moving average interpolation about how large the window should be. Clearly, if the distance separating an unvisited site from a data point is greater than the range, then that data point can make no useful contribution to the interpolation; it is too far away.

The third point shown by Figure 6.2 is that the fitted model does not pass through the origin, but cuts the y-axis at a positive value of $\gamma(\mathbf{h})$. According to equation 6.22 the semivariance is zero when $\mathbf{h} = 0$, because the differences between points and themselves is by definition zero. The positive value of $\gamma(\mathbf{h})\ \mathbf{h} \to 0$ is an estimate of ε'', the residual, spatially uncorrelated noise. ε'' is known as the *nugget*; this is the variance of

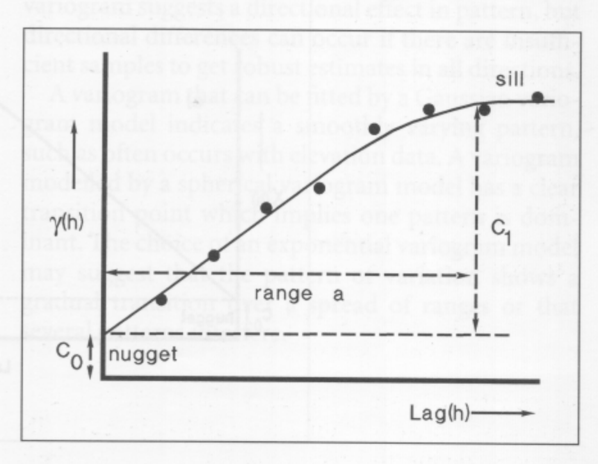


Figure 6.2. An example of a simple transitional variogram with range, nugget, and sill

measurement errors combined with that from spatial variation at distances much shorter than the sample spacing, which cannot be resolved.

The form of the variogram can be quite revealing about the kind of spatial variation present in an area, and can help to decide how to proceed further. When the nugget variance is important but not too large, and there is a clear range and sill, a curve known as the spherical model,

$$\begin{aligned}\gamma(\mathbf{h}) &= c_0 + c_1\left\{\frac{3\mathbf{h}}{2a} - 1/2(\mathbf{h}/a)^3\right\} && \text{for } 0 < \mathbf{h} < a \\ &= c_0 + c_1 && \text{for } \mathbf{h} \geq a \\ \gamma(0) &= 0 && \end{aligned} \qquad 6.6$$

where a is the range, **h** is the lag, c_0 is the nugget variance, and $c_0 + c_1$ equals the sill, often fits observed variograms well.

If there is a clear nugget and sill, but only a gradual approach to the range, the exponential model

$$\gamma(\mathbf{h}) = c_0 + c_1\,\{1 - exp\,(-\mathbf{h}/a)\} \qquad 6.7$$

is often a good choice.

If the variation is very smooth and the nugget variance ε'' is very small compared to the spatially dependent random variation $\varepsilon''(\mathbf{x})$, then the variogram can often best be fitted by a curve having an inflection, such as the Gaussian model.

$$\gamma(\mathbf{h}) = c_0 + c_1\,\{1 - exp\,(-\mathbf{h}^2/a^2)\} \qquad 6.8$$

All these models are known as *transitive variograms* because the spatial correlation structure varies with **h**; *non-transitive variograms* have no sill within the area sampled and may be modelled by the linear model:

$$\gamma(\mathbf{h}) = c_0 + bh \qquad 6.9$$

where b is the slope of the line. A linear variogram typifies attributes which vary at all scales, such as simple Brownian motion. Figure 6.3 shows examples of these variograms. A variogram that becomes increasingly

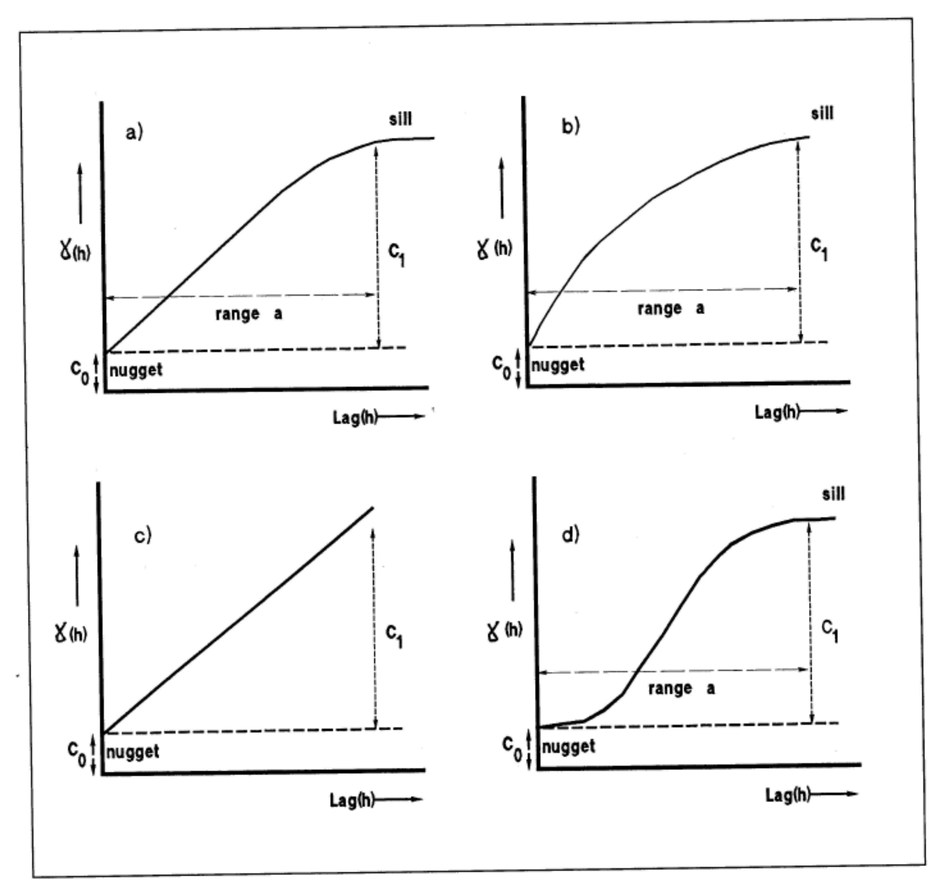


Figure 6.3. Examples of the most commonly used variogram models: (*a*) spherical; (*b*) exponential; (*c*) linear; and (*d*) Gaussian

steep with **h** indicates a trend in the data that should be modelled separately.

Variogram estimation and modelling is extremely important for structural analysis and for interpolation. The variogram models cannot be any haphazardly chosen function, as they must obey certain mathematical constraints (see Isaaks and Srivastava 1989). Also the models must not be fitted to the experimentally estimated semivariances for the various lags without taking the numbers of pairs of points at each lag into consideration. Because the data are used repeatedly when estimating variograms, the effective degrees of freedom are greatest for the shortest lag and then decrease in a complex way (cf. Taylor and Burrough 1986) with the lag. Consequently the fitting of variogram models usually proceeds using a weighted least squares method where the weights are computed from the numbers of pairs. Even so, variogram fitting is an interactive process requiring considerable judgement and skill. Modern Geostatistics software usually includes interactive routines for variogram fitting (e.g. Pebesma 1996, Pannatier 1996).

Using the variogram for spatial analysis

The variogram is an essential step on the way to determining optimal weights for interpolation. When the nugget variance ε'' so dominates the local variation that the experimental variogram shows no tendency to diminish as $\mathbf{h} \to 0$, the interpretation is that the data are so noisy that interpolation is not sensible. In this situation, the best estimate of $z(\mathbf{x})$ is the overall mean computed from all sample points in the region of interest without taking spatial dependence into account: interpolation is meaningless and a waste of time and money.

A noisy variogram, in which the experimentally derived semivariances are scattered, suggests that too few samples have been used to compute $\hat{\gamma}(\mathbf{h})$. A rule of thumb suggests that possibly at least 50–100 data points are necessary to achieve a stable variogram, depending on the kind of spatial variation encountered, though smooth surfaces require fewer points than those with irregular variation: smoother variograms can also be obtained by increasing the size of the search window.

The range of the variogram provides clear information about the size of the search window that should be used. If the distance from a data point to an unsampled point exceeds the range, then it is too far away to make any contribution; if all data points are further away than the range, the best estimate is the general mean. These distances can be modified by anisotropy, which modifies the shape of the search neighbourhood from a circle to an ellipse.

The presence of a *hole effect* in the experimental variogram (a dip in semivariance at distances greater than the range) may indicate a pseudo-periodic pattern caused by long-range variation over a study area that is too small to encompass the total range of variation. True periodicity will give a variogram with a periodic variation in the sill that matches the wavelength of the pattern, providing the original field sampling is in step with the periodicity.

If the range is large then long-range variation dominates: if it is small, then the major variation occurs over short distances. Anisotropy in the experimental variogram suggests a directional effect in pattern, but directional differences can occur if there are insufficient samples to get robust estimates in all directions.

A variogram that can be fitted by a Gaussian variogram model indicates a smoothly varying pattern, such as often occurs with elevation data. A variogram modelled by a spherical variogram model has a clear transition point which implies one pattern is dominant. The choice of an exponential variogram model may suggest that the pattern of variation shows a gradual transition over a spread of ranges or that several patterns interfere.

Isotropic and anisotropic variation

In the foregoing, we have implied that the source data for the variogram are collected on regularly spaced transects, or possibly a regular grid. In many cases we only have irregularly spaced measurements, so it is useful to be able to compute the experimental variogram from such data. To do this we use a circular search radius, rather like the tyre of a bicycle, to define a zone whose midpoint is **h** from its centre. This wheel is placed over a data point and all data points falling in the tyre are used to estimate the contribution of $(z_i - z_j)^2$ from all pairs (Figure 6.4). As a general rule of thumb to avoid edge effects, it is not usually sensible to compute a variogram for more lags than a total separation distance of one-half the dimensions of the study area. If we ignore directional effects, the resulting variogram is known as *isotropic*, it averages the variogram over all directions. However, as Figure 6.4 shows, it is easy to compute the variogram for specific directions β. These are known as *anisotropic* variograms, and if different in range or sill for different values of β they may indicate that the spatial variation varies with direction. This could happen in sediments perpendicular or parallel to a river, for example.

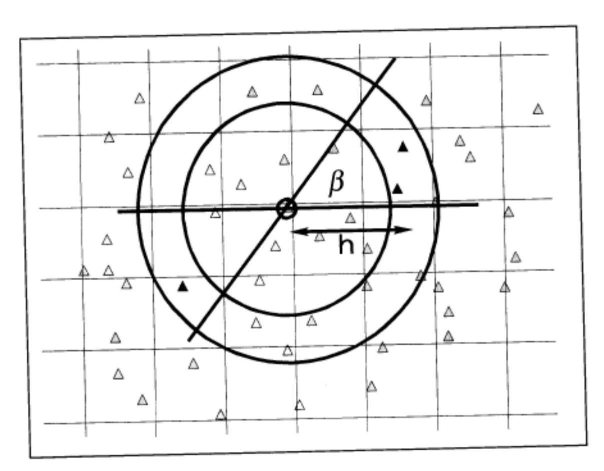


Figure 6.4. Circular search window for estimating semivariance in any direction ß

By computing $(z_i - z_j)^2$ for all possible directions one creates a variogram surface, which can be plotted as an ellipsoid map with centre point 0,0. The *major axis* of the variogram is the longer axis of the ellipse, and its orientation is given by the angle of that axis with respect to north.

Variograms showing spatial variation at several scales

The original model of covariance simply assumes that the spatial variation of the attribute of interest can be divided into the three components modelled by (i) a general mean or trend, (ii) a variogram, and (iii) residual nugget. In some situations, particularly if there is sufficient data, it may be possible and useful to distinguish more than one variogram component, for example where two random patterns having two or more widely separated ranges interfere with each other. In this case, the complex variogram $\gamma_T(\mathbf{h})$ can be split into several components according to the linear model of coregionalization (Isaaks and Srivastava 1989):

$$\gamma_T(\mathbf{h}) = \gamma_1(\mathbf{h}) + \gamma_2(\mathbf{h}) + \gamma_3(\mathbf{h}) + \dots \qquad 6.10$$

so for two levels of nesting we replace equation 6.4 by:

$$Z(\mathbf{x}) = m(\mathbf{x}) + \gamma_1(\mathbf{h}) + \gamma_2(\mathbf{h}) + \varepsilon'' \qquad 6.11$$

Each of the sub-variograms is defined by its own set of parameters.

Spatial variation within different cover classes

The straightforward approach to computing the variogram is to assume that all data are located in the same cover class or domain. In this case the computed variogram is a global model that determines the interpolation weights over the whole area. This need not necessarily be so: different parts of an area may have important differences in land cover, rock type, soil type, flooding frequency, or land tenure pattern, each of which has its own, unique spatial correlation structure or pattern. For example, the pattern of distribution of heavy metals in river floodplains is likely to be quite different from that on the hillsides bordering the river, because the sources of heavy metals and the transport processes will be quite different.

When faced with the problem of important differences in domain it may be sensible to see if the study area can be better modelled by a set of domain-specific variograms rather than a single global model. The main problem is usually insufficient data: should one risk using an inappropriate, but well-defined global variogram, or several poorly defined, local models? The pragmatic approach is to attempt to use local models whenever possible because then for each spatial unit (polygon or mapping unit) the variogram for any given variable, becomes a new, unique attribute of that polygon, and its parameters can be stored in the relational database (Pebesma 1996, Laurini and Pariente 1996, in Burrough and Frank 1996. Voltz and Webster 1990).

Using the variogram for interpolation: ordinary kriging

Given that the spatially dependent random variations are not swamped by uncorrelated noise, the fitted variogram can be used to determine the weights λ_i needed for local interpolation. The procedure is similar to that used in weighted moving average interpolation except that now the weights are derived from a geostatistical analysis of the data rather than from a general, and possibly inappropriate, model. We have:

$$\hat{z}(\mathbf{x}_0) = \sum_{i=1}^{n} \lambda_i \cdot z(\mathbf{x}_i) \qquad 6.12$$

with $\sum_{i=1}^{n} \lambda_i = 1$. The weights λ_i are chosen so that the estimate $\hat{z}(\mathbf{x}_0)$ is unbiased, and that the estimation variance σ_e^2 is less than for any other linear combination of the observed values.

The minimum variance of $\hat{z}(\mathbf{x}_0)$ is given by:

$$\hat{\sigma}_e^2 = \sum_{i=1}^{n} \lambda_i \gamma(x_i, x_0) + \phi \qquad 6.13$$

and is obtained when

$$\sum_{i=1}^{n} \lambda_i \gamma(x_i, x_j) + \phi = \gamma(x_j, x_0) \quad \text{for all } j \qquad 6.14$$

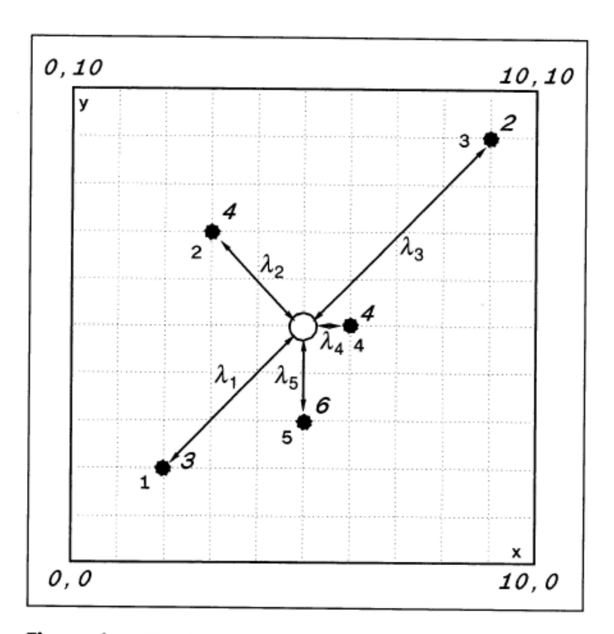


Figure 6.5. Simple example to illustrate the prediction of *z* at unsampled sites by ordinary kriging

BOX 6.2. A WORKED EXAMPLE OF COMPUTING KRIGING WEIGHTS

Computing kriging weights for the unsampled point $z(x_{i=0})$ ($x = 5, y = 5$) in Figure 6.5

Let the spatial variation of the attribute sampled at the five points be modelled by a spherical variogram with parameters $c_0 = 2.5$, $c_1 = 7.5$ and range $a = 10.0$. The data at the five sampled points are:

i	*x*	*y*	*z*
1	2	2	3
2	3	7	4
3	9	9	2
4	6	5	4
5	5	3	6

In matrix terms we have to solve the following:

$$A^{-1} \cdot b = \begin{bmatrix} \lambda \\ \phi \end{bmatrix}$$

where *A* is the matrix of semivariances between pairs of data points, *b* is the vector of semivariances between each data point and the point to be predicted and λ is the vector of weights. ϕ is a Lagrangian for solving the equations.

Start by creating a distance matrix between the data points:

i	1	2	3	4	5
1	0.0	5.099	9.899	5.000	3.162
2	5.099	0.0	6.325	3.606	4.472
3	9.899	6.325	0.0	5.0	7.211
4	5.0	3.606	5.0	0.0	2.236
5	3.162	4.472	7.211	2.236	0.0

and the vector of distances between the data points and the unknown site:

i	0
1	4.243
2	2.828
3	5.657
4	1.0
5	2.0

Substitute these numbers into the variogram to get the corresponding semivariances (matrices *A* and *b*).

A = *i*	1	2	3	4	5	6
1	2.500	7.739	9.999	7.656	5.939	1.000
2	7.739	2.500	8.667	6.381	7.196	1.000
3	9.999	8.667	2.500	7.656	9.206	1.000
4	7.656	6.381	7.656	2.500	4.936	1.000
5	5.939	7.196	9.206	4.936	2.500	1.000
6	1.000	1.000	1.000	1.000	1.000	0.000

Note the extra row and column ($i = 6$) to ensure that the weights sum to 1.

b = *i*	0
1	7.151
2	5.597
3	8.815
4	3.621
5	4.720
6	1.000

$A^{-1} = i$

i	1	2	3	4	5	6
1	−.172	.050	.022	−.026	.126	.273
2	.050	−.167	.032	.077	.007	.207
3	.022	.032	−.111	.066	−.010	.357
4	−.026	.077	.066	−.307	.190	.030
5	.126	.007	−.010	.190	−.313	.134
6	.273	.207	.357	.003	.134	−6.873

with λ as:

i	weights		distances
1	.0175		4.423
2	.2281		2.828
3	−.0891	$\Sigma = 1$	5.657
4	.6437		1.000
5	.1998		2.000
6	.1182	ϕ	

Therefore the value at $z(x_{i=0}) = .0175 * 3 + .2281 * 4 - .0891 * 2 + .6437 * 4 + .1998 * 6$
$= \mathbf{4.560}$

with estimation variance $\sigma_e^2 = [.0175 * 7.151 + .2281 * 5.597 - .0891 * 8.815$
$+ .6437 * 3.621 + .1998 * 4.720] + \phi$
$= 3.890 + 0.1182$
$= \mathbf{4.008}$

The quantity $\gamma(\mathbf{x}_i,\mathbf{x}_j)$ is the semivariance of z between the sampling points $\mathbf{x}_i$ and $\mathbf{x}_j$; $\gamma(\mathbf{x}_i,\mathbf{x}_0)$ is the semivariance between the sampling point $\mathbf{x}_i$ and the unvisited point $\mathbf{x}_0$. Both these quantities are obtained from the fitted variogram. The quantity ϕ is a Lagrange multiplier required for the minimalization. The method is known as *ordinary kriging*: it is an exact interpolator in the sense that when the equations given above are used, the interpolated values, or best local average, will coincide with the values at the data points. In mapping, values will be interpolated for points on a regular grid that is finer than the spacing used for sampling, if that was itself a grid. The interpolated values can then be converted to a contour map using the techniques already described. Similarly, the estimation error σ_e^2, known as the *kriging variance*, can also be mapped to give valuable information about the reliability of the interpolated values over the area of interest. Often the kriging variance is mapped as the *kriging standard deviation* (or kriging error), because this has the same units as the predictions, and this convention is followed in this chapter. Box 6.2 and Figure 6.5 show how the equations for ordinary kriging are set up and solved.

Example: Figure 6.6 presents the results of using ordinary kriging for the untransformed zinc data. The map of prediction variances is also shown. Compare the prediction map with the inverse distance maps and spline map (Figure 5.10) and note that it is smoother than these, but also avoids the local large values that are linked to isolated data points. The standard deviation map clearly shows how prediction variance is linked to the density of the data points. Though comparisons cannot strictly be made because variance analysis and kriging use different statistical models (see de Gruijter and Ter Braak 1990), note that the maximum kriging standard deviation of 458 ppm is similar to the maximum within-class standard deviation of class 1 of the flooding frequency map (Figure 5.2*c*), but that this maximum only occurs in the kriging map in areas where there are no data points. In most of the areas covered by class 1 (standard deviation 423 ppm) the kriging prediction standard error is about 140 ppm. These results suggest that kriging prediction that takes account of gradual spatial change appears to be much better than that given by global, crisp, choropleth mapping.

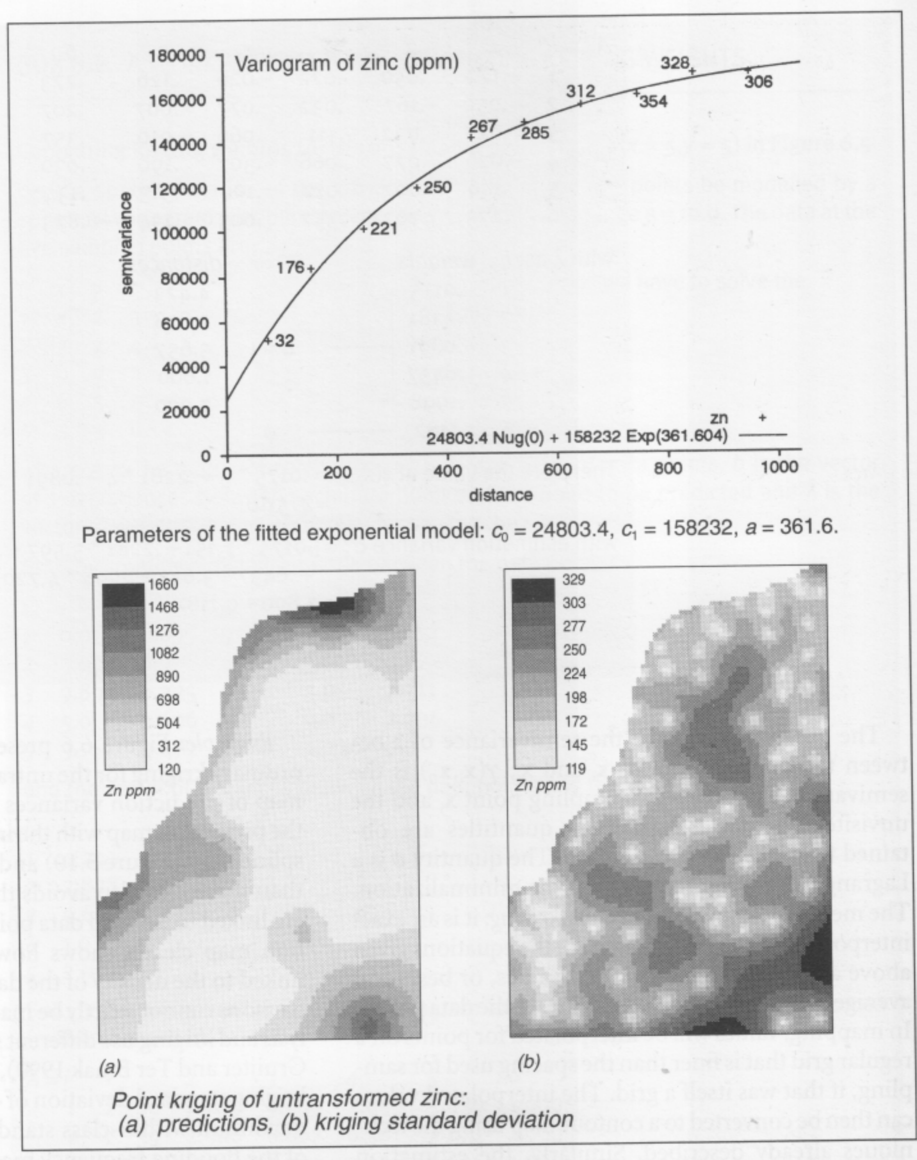


Figure 6.6. Results of interpolating the Maas data set using ordinary kriging. Above—variogram with fitted exponential model; (*a*) Kriging predictions of zinc levels, (*b*) Kriging standard deviations

Using kriging to validate the variogram model

Cross validation is the practice of using the kriging equations retrospectively to check the variogram model. It involves computing the moments of the distribution of $(\hat{z}(x_i) - z(x_i))$ for all data points, when each data point is successively left out and predicted from the rest of the data (Figure 6.7). The procedure is designed only to test the variogram for self-consistency and lack of bias, indicated by a mean difference near zero and a *z-score* (not to be confused with the *z* data values!) of one.

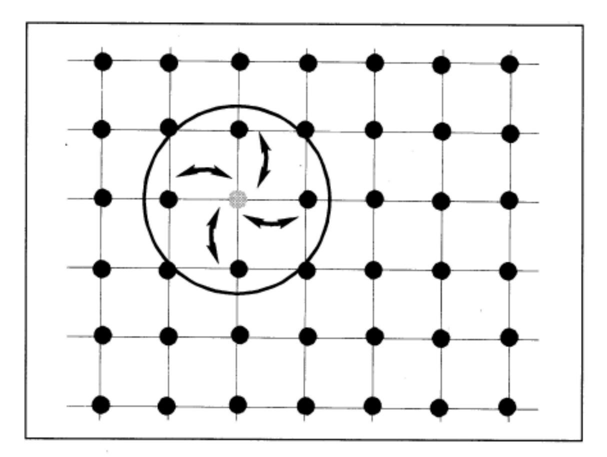


Figure 6.7. Cross-validation is the process of checking the variogram against the original data

Block kriging

Clearly, kriging fulfils the aims of finding better ways to estimate interpolation weights and of providing information about errors. The resulting map of interpolated values may not be exactly what is desired, however, because the point kriging, or simple kriging, equations (6.13 and 6.14) imply that all interpolated values relate to the *support*, which is the area or volume of an original sample. Very often, as in sampling for soil or water quality, this sample is only a few centimetres across. Given the often large amplitude, short-range variation of many natural phenomena like soil or water quality (Burrough 1993), ordinary point kriging may result in maps that have many sharp spikes or pits at the data points. This can be overcome by modifying the kriging equations to estimate an average value $z(B)$ of the variable z over a block of land B. This is useful when one wishes to estimate average values of z for experimental plots of a given area, or to interpolate values for grid cells of a specific size for quantitative modelling (Figure 6.8).

The average value of z over a block B, given by

$$z(B) = \int_B \frac{z(\mathbf{x})\mathrm{d}\mathbf{x}}{\mathrm{area}B} \qquad 6.15$$

is estimated by

$$\hat{z}(B) = \sum_{i=1}^{n} \lambda_i \cdot z(\mathbf{x}_i) \qquad 6.16$$

with $\sum_{i=1}^{n} \lambda_i = 1$, as before.

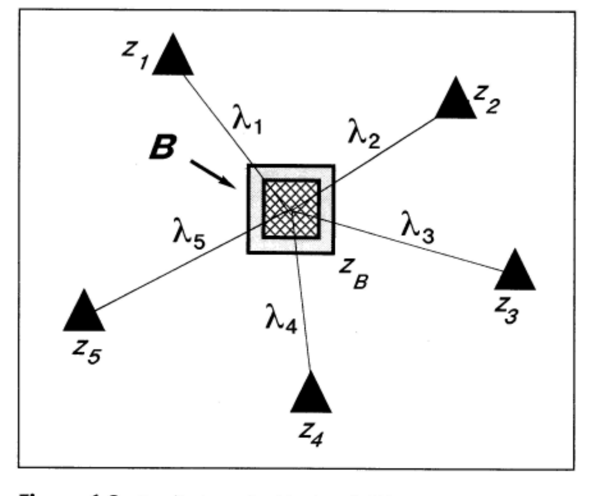


Figure 6.8. Predicting z for blocks of different size

The minimum variance is now

$$\hat{\sigma}^2(\boldsymbol{B}) = \sum_{i=1}^{n} \lambda_i \bar{\gamma}(\boldsymbol{x}_i, \boldsymbol{B}) + \phi - \bar{\gamma}(\boldsymbol{B}, \boldsymbol{B}) \qquad 6.17$$

and is obtained when

$$\sum_{i=1}^{n} \lambda_i \gamma(\boldsymbol{x}_i, \boldsymbol{x}_j) + \phi = \bar{\gamma}(\boldsymbol{x}_j, \boldsymbol{B}) \quad \text{for all } j \qquad 6.18$$

Estimation variances obtained for block kriging are usually substantially less than for point kriging. When these equations are used, the resulting smoothed interpolated surface is free from the pits and spikes resulting from point kriging.

Other forms of kriging

Simple kriging

Simple kriging is prediction by generalized linear regression under the assumption of second order stationarity with a known mean (Olea 1991). Because the assumption of second order stationarity is often too restrictive for most data from the physical environment, ordinary kriging (no a priori mean) is most often used, though Voltz and Webster (1990) and Webster and McBratney (1987) provide examples.

Non-linear kriging

Lognormal kriging is the interpolation of log-normally distributed, rather than normally distributed data (Deutsch and Journel 1992). The data are first transformed to natural logarithms or base-10 logarithms so that variogram modelling and interpolation proceeds with the transform $y(\mathbf{u}) = \ln z(\mathbf{u})$ to give an estimate $\hat{y}(\mathbf{u})$ for $\ln z(\mathbf{u})$. The predicted values can be transformed back after interpolation but care must be exercised because the antilog back-transform $e^{y^*(u)}$ is a biased estimator of $Z(\mathbf{u})$. Deutsch and Journel recommend using an unbiased back-transform

$$z^*(\mathbf{u}) = \exp\left[y^*(u) + \frac{\sigma_{SK}^2(u)}{2}\right] \qquad 6.19$$

where $\sigma_{SK}^2(\mathbf{u})$ is the simple lognormal kriging variance. Deutsch and Journel (1992) point out that the extreme sensitivity of the errors for antilog back-transformation make lognormal kriging difficult to use. They recommend Multigaussian kriging (MG) or Indicator Kriging (see section 6.5 below) as alternatives. Nevertheless, the lognormal transform is useful for many physical data with positive skew distributions such as soil chemical and physical attributes (Webster and Oliver 1990, Burrough *et al.* 1992).

Figure 6.9*a* presents the experimental variogram of the zinc data using the log transformed data and Figure 6.10*a,b* show the maps of predictions and standard deviations of the log-transforms. The general patterns are similar to those in Figure 6.6.

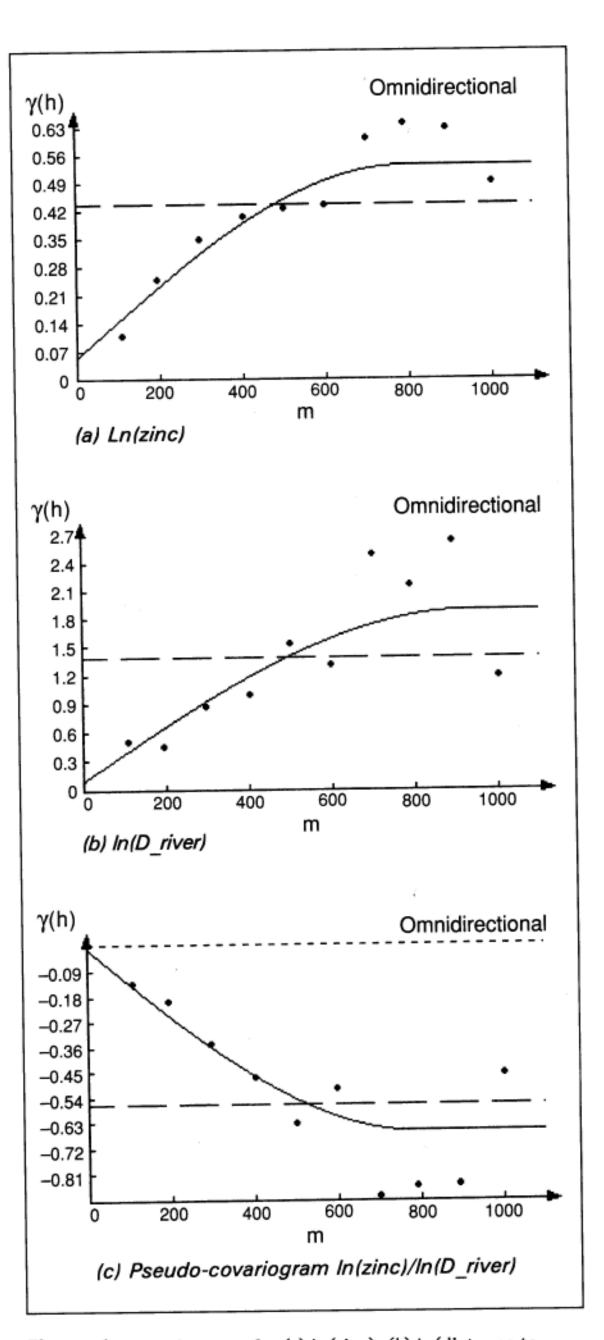


Figure 6.9. Variograms for (*a*) ln(zinc), (*b*) ln(distance to river), and (*c*) pseudo cross variogram

Ordinary kriging with anisotropy or nested variograms

Incorporating anisotropy into the ordinary kriging procedure is simply a matter of modifying the con-

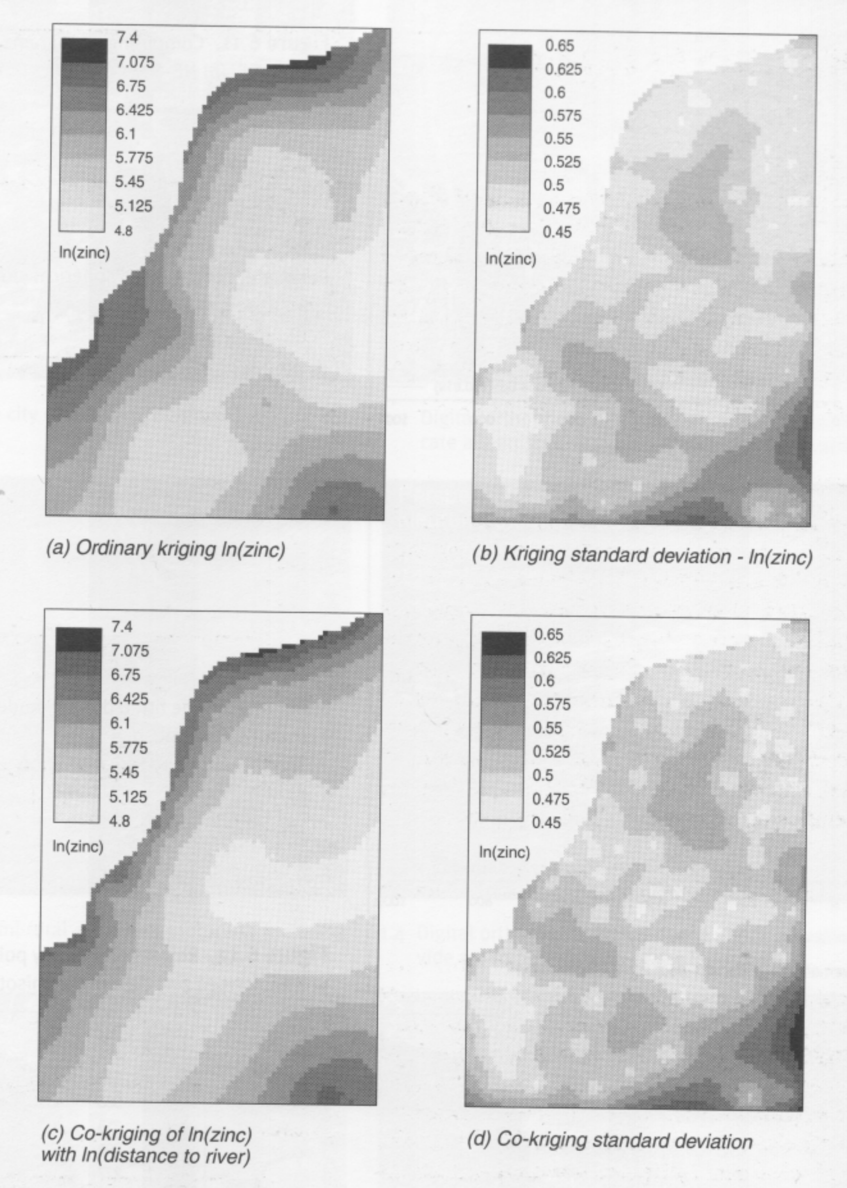
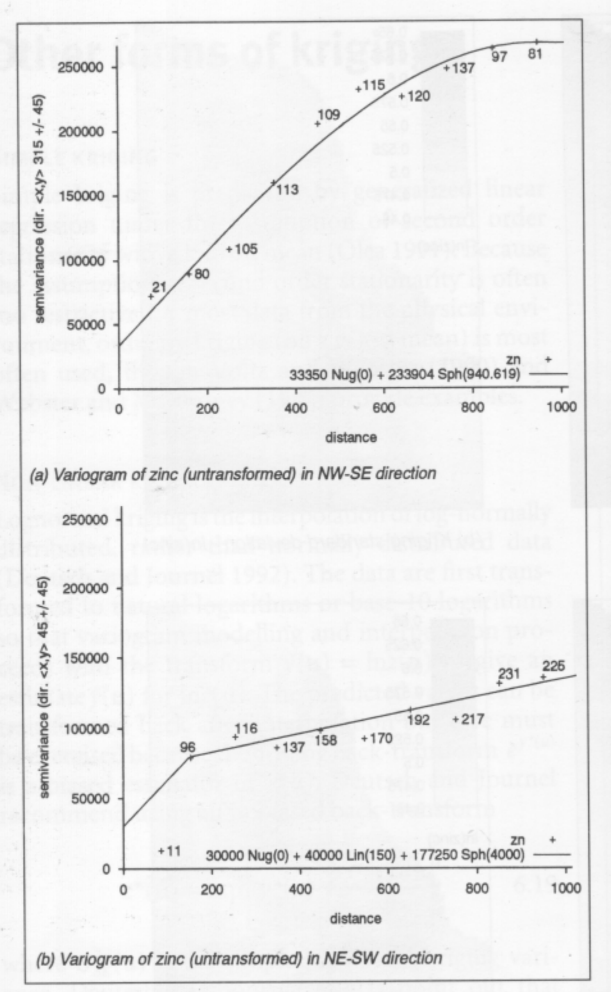


Figure 6.10. A comparison of ordinary point kriging and co-kriging using log-transformed data

version of the distance matrix into the matrix of semivariances *A* (see Box 6.2) taking account the variation of semivariance with direction. Figure 6.11 presents the variograms computed in a NW–SE direction (perpendicular to the river) and in a NE–SW direction (parallel to the river). Note the double variogram fitted to the major axis (NE–SW), which is an attempt to force both variograms to have the same nugget. Figure 6.12 shows the effect of incorporating anisotropy into the interpolation, producing long, linear streaks parallel to the river. The results are plausible, but in spite of the clear variograms they look rather contrived compared to the other maps.



(a) Variogram of zinc (untransformed) in NW-SE direction

(b) Variogram of zinc (untransformed) in NE-SW direction

Figure 6.11. Computing variograms in different directions (*a*) NW–SE, (*b*) NE–SW

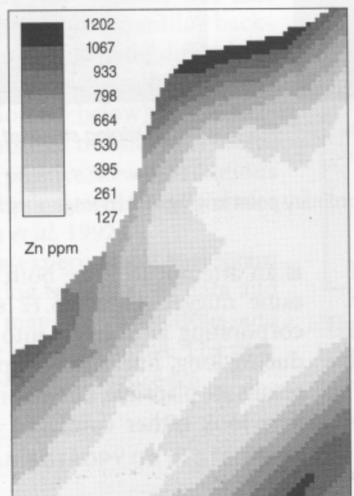
Parameters of the fitted spherical model: $c_0 = 33350$, $c_1 = 233909$, $a = 940.6$.

Parameters of the fitted double model: $c_0 = 30000$

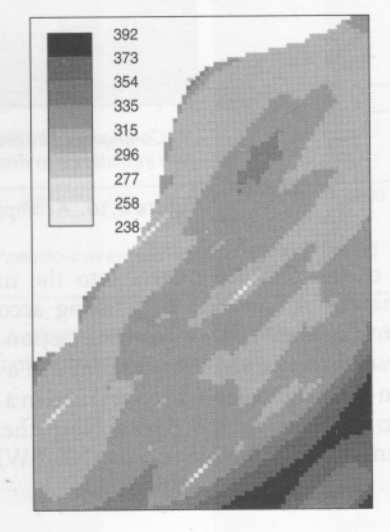
Linear $c_1 = 40000$, $a_1 = 150$

Spherical $c_2 = 177250$, $a_2 = 4000$

Figure 6.12. Results of ordinary point kriging of untransformed zinc data using anisotropic variograms



(a) Zinc levels from anisotropic variogram



(b) Anisotropic kriging standard deviation

Kriging using extra information

Frequently, the data points are not the only source of information about the distribution of z, and we may be able to draw on other knowledge that can help with interpolation. The main sources are: (*a*) an appropriate stratification into clearly different domains, (*b*) data from a cheap-to-measure co-variable that has been sampled at many more data points, and (*c*) a physical or empirical spatial model of a driving process.

STRATIFIED KRIGING

When there is enough soft information to classify the area into meaningful sub-areas and there are enough data to compute variograms for each different domain, the interpolation can be carried out on points or blocks for each area separately, adjusting the kriging equations to avoid discontinuities at class boundaries. The analysis of variance of zinc concentration according to flood frequency classes (Chapter 5; Figure 5.2) suggested that there were two classes of importance: flood class 1 (annual flooding) and the rest. Figure 6.13 presents the variograms computed for these two classes. It shows that both units have strongly correlated spatial variation, but that the variance in flood class 1 is almost ten times as great as in the less frequently flooded areas. Figure 6.14 shows the interpolations which show how the zinc concentrations vary within the classes. Comparing the stratified kriging standard deviation map with Figure 6.6*a*,*b* (ordinary point kriging) demonstrates how the stratification reduces the interpolation errors, and for much of the area reduces their dependence on the distribution of data points. In this case, stratification also preserves the fine spatial structure of the narrow, flooded channels, which is smaller that the average spacing between data points and is therefore lost by interpolating from the 'hard' point data alone.

CO-KRIGING

Often data may be available for more than one attribute per sampled location. One set (*U*) may be expensive to measure and therefore is sampled infrequently while another (*V*) may be cheap to measure and has more observations. If *U* and *V* are spatially correlated then it may be possible to use the information about the spatial variation of *V* to help map *U*. This can be done by using an extension of the normal kriging technique known as *co-kriging*. Co-kriging can produce predictions for both points and blocks in analogy to ordinary kriging (Isaaks and Srivastava 1989).

Besides the variograms describing the non-structural variation of both U and V, co-kriging needs information on the *joint* spatial covariation of both

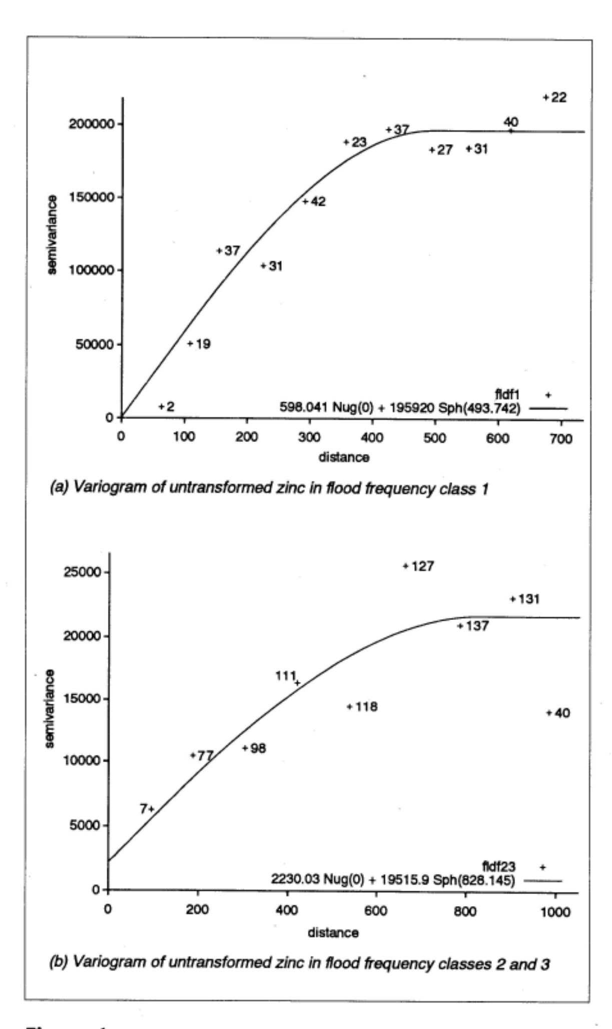
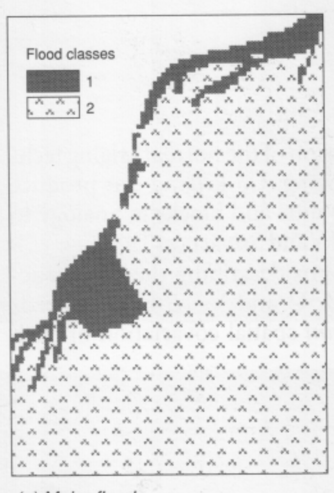
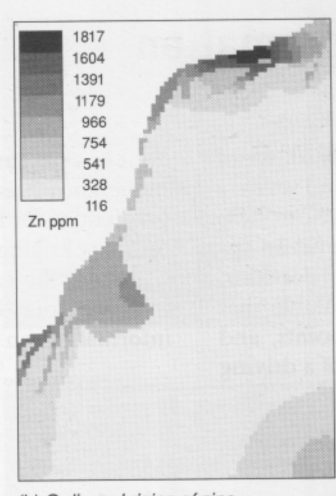


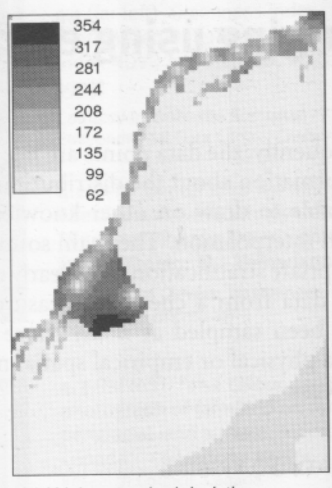
Figure 6.13. Computing the variograms of zinc for flood frequency zones separately. (*a*) variogram for flood frequency class 1; (*b*) variogram for flood frequency classes 2 and 3. Note that the sill for class 1 is nearly 10 times higher than for classes 2 and 3



(a) Major flood divisions



(b) Ordinary kriging of zinc within flood classes (ppm)



(c) Kriging standard deviation within flood classes

Figure 6.14. Prediction of zinc levels using stratified variograms in Figure 6.13

variables. For any pair of variables U and V the cross-semivariance $\gamma_{UV}(\boldsymbol{h})$ at lag $\boldsymbol{h}$ is defined as:

$$2\gamma_{UV}(\boldsymbol{h}) = E[\{Z_U(x) - Z_U(x + \boldsymbol{h})\}\{Z_V(x) - Z_V(x + \boldsymbol{h})\}] \quad 6.20$$

where Z_U, Z_V are the values of U,V at places x, $x + \boldsymbol{h}$.

The *cross-variogram* is estimated directly from the sample data using:

$$\hat{\gamma}_{UV}(\boldsymbol{h}) = 1/2n(\boldsymbol{h})\sum_{i=1}^{n(\boldsymbol{h})}\{z_U(x_i) - z_U(x_i + \boldsymbol{h})\}\{z_V(x_i) - z_V(x_i + \boldsymbol{h})\} \quad 6.21$$

where $n(\boldsymbol{h})$ is the number of data pairs of observations of z_U, z_V at locations x_i, $x_i + \boldsymbol{h}$ for the distance vector $\boldsymbol{h}$. Cross-variograms can increase or decrease with $\boldsymbol{h}$ depending on the correlation between U and V. When cross-variograms are fitted, the Cauchy-Schwartz relation:

$$|\gamma_{UV}(\boldsymbol{h})| \leq \sqrt{(\gamma_U(\boldsymbol{h})^*\gamma_V(\boldsymbol{h}))} \quad \text{for all } \boldsymbol{h} > 0 \quad 6.22$$

must be checked to guarantee a positive co-kriging estimation variance in all circumstances (Deutsch and Journel 1992; Isaaks and Srivastava 1989).

A co-kriged estimate is a weighted average in which the value of U at location x_0 is estimated as a linear weighted sum of co-variables V_k. If there are k variables $k = 1, 2, 3, \ldots V$ and each is measured at n_V places, $x_{ik} = 1, 2, 3, \ldots N_k$, then the value of one variable U at x_0 is predicted by:

$$\hat{z}_U(x_0) = \sum_{k=1}^{V}\sum_{i=1}^{n_V}\lambda_{ik} z(x_{ik}) \quad \text{for all } V_k \quad 6.23$$

To avoid bias, i.e. to ensure that $E[z_U(x_0) - \hat{z}_U(x_0)] = 0$ the weights λ_{ik} must sum as follows:

$$\sum_{i=1}^{n_V}\lambda_{ik} = 1 \quad \text{for } V = U \quad \text{and}$$

$$\sum_{i=1}^{n_V}\lambda_{ik} = 0 \quad \text{for all } V_k \neq U \quad 6.24$$

The first condition implies that there must be at least one observation of U for co-kriging to be possible. The interpolation weights are chosen to minimize the variance:

$$\sigma_U^2(x_0) = E[\{z_U(x_0) - \hat{z}_U(x_0)\}^2] \quad 6.25$$

There is one equation for each combination of sampling site and attribute, so for estimating the value of variable j at site x_0 the equation for the g-th observation site of the k-th variable is:

$$\sum_{j=1}^{V}\sum_{i=1}^{n_V}\lambda_{ij}\gamma_{ij}(x_{ij},x_{gk}) + \phi_k = \gamma_{UV}(x_0,x_{gk}) \quad 6.26$$

for all $g = 1$ to n_V and all $k = 1$ to V, where ϕ_k is a Lagrange multiplier. These equations together make up the co-kriging system.

Co-kriging only has an advantage if the cheap-to-measure attributes and the expensive attribute are measured at different numbers of data points. If both (or more) attributes are all measured at the same points then the co-kriging will not give estimates that are different from ordinary kriging.

If the cross-variogram is computed between variables measured at the same set of observation points it is called a true cross-variogram. If it is computed for pairs of attributes measured on observations at different locations then it is known as the *pseudo cross-variogram* because an extra assumption has been made, namely that both sets of sample points belong to comparable realizations of the regionalized variables.

Example. In Chapter 5, equation 5.9 and Figure 5.5 we noted a strong, linear correlation between ln(zinc) levels and ln(distance to river). For illustration, assume that we have 98 measurements of 'Distance to river' but only 49 of zinc. Figure 6.9 shows the variograms for ln(zinc) and ln(Distance to river) and the cross variogram between both variables. Figure 6.20*c*,*d* shows the point co-kriging of ln(zinc) based on the 49 observations for zinc aided by the 98 of ln(Distance to river). Comparison of Figure 6.10*d* with Figure 6.10*b* (standard deviation for ln(zinc) based on 98 data points) shows that the interpolation error of the co-kriged map (on the logarithmic scale) is approximately the same as that of the point kriging of ln(zinc) alone. This illustrates the potential of co-kriging to save on expensive laboratory analyses if there is strong spatial correlation with a cheap-to-measure co-variable.

Many examples of co-kriging in soil mapping are listed in Burrough 1993; Solow and Gorelick (1986) give an example for estimating streamflow. Co-kriging is most successful when the patterns of the variables used are related by a common physical process.

UNIVERSAL KRIGING—INTERPOLATION WITH A BUILT-IN TREND

Co-kriging is a difficult technique to carry out, requiring considerable insight into the theory of geostatistics and some authorities feel that the effort in fitting cross-variograms that obey the limitations of the intrinsic hypothesis and the linear model of co-regionalization is not worthwhile. An alternative that appeals to many environmental scientists is to use the knowledge from empirical (regression) transfer models as part of the kriging procedure. This is sometimes known as *Universal* or *KT Kriging* (Deutsch and Journel 1992) or kriging in the presence of external trends. Modern theory has completely supplanted the first ideas of universal kriging reviewed in Burrough (1986). Isaacs and Srivastava (1989) point out that the term $m(\mathbf{x})$ in equation (6.1) where regionalized variable Z at $\mathbf{x}$ given by

$$Z(\mathbf{x}) = m(\mathbf{x}) + \varepsilon'(\mathbf{x}) + \varepsilon'' \qquad 6.27$$

can be modelled by a deterministic trend, such as a regression equation that is incorporated in the kriging equations.

$$Z(\mathbf{x}) = \sum_{j=1}^{p} f_j(\mathbf{x})\beta_j + \gamma(\mathbf{x}) \qquad 6.28$$

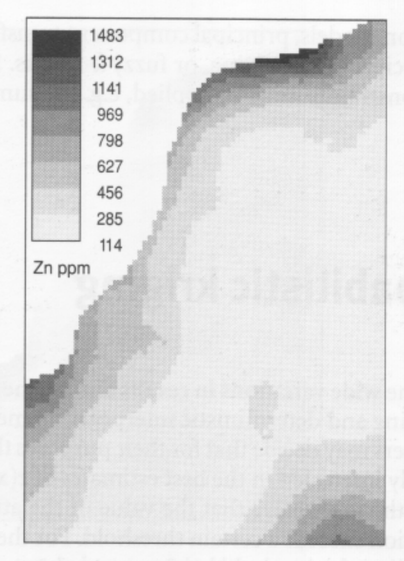


Figure 6.15. Prediction of zinc levels with ordinary kriging incorporating a trend model (predictions made with log transformed data, back transformed for display)

Figure 6.15 shows the result of computing zinc concentrations by universal kriging in a procedure that combines the regression modelling of ln(zinc) on elevation and ln(Distance to river). A single variogram for ln(zinc) was used for the whole area. Compare this result with Figure 5.6 (simple regression) and 6.10*a* (ordinary kriging).

MULTIVARIATE KRIGING

Multivariate kriging is the application of geostatistics to multivariate transformations, such as the results of

regression models, principal component transformations, reciprocal averaging, or fuzzy k-means. Sometimes constraints must be applied, e.g. the sum of all fuzzy k-means values must equal 1, so the kriging equations need to be modified. See Chapter 11 and de Gruijter *et al.* (1997).

Probabilistic kriging

Given the wide variations in results among the different kriging and deterministic interpolation methods, some users may decide that for their purposes they are not really interested in the best estimates of $z(\mathbf{x}_0)$, but only in the *probability* that the value of the attribute in question exceeds a certain threshold. For them, the probability of a threshold being exceeded may be sufficient to support a decision to mine, to commission operations to clean up polluted soil or to set up a marketing base. *Indicator kriging* is a non-linear form of ordinary kriging in which the original data are transformed from a continuous scale to a binary scale (1 if $z \leq T_j$, and 0 otherwise, or vice-versa). Variograms are computed for the binary data in the usual way, and ordinary kriging proceeds with the transformed data. The resulting maps display continuous data in the range 0–1 indicating the probability that T_j has been exceeded or not exceeded, as the case may be. Computing variograms and interpolating for other thresholds provides insight in how the probabilities of threshold exceedance vary with threshold levels.

Figure 6.16 presents indicator variograms for cut-offs in zinc levels at 500 ppm and 1000 ppm measured on all 98 points. The resulting maps (Figures 6.17*a*,*b*) show the *probability that the zinc level exceeds the given threshold.* The results are clear and easy to understand and could easily be incorporated in a GIS database for decision-making where the managers do not want to have to understand the intricacies of kriging.

Indicator kriging can easily be combined with soft information, such as the flood frequency classes: Figure 6.17*c*,*d* show the probabilities of zinc levels exceeding 500 ppm within the areas defined by the flood frequency classes, and these figures provide some interesting insights into the process of flooding and contamination around frequently flooded creeks.

There are several published examples of indicator kriging and indicator co-kriging in soil science, soil pollution, and geology (Journel and Posa 1990, Okx and Kuipers 1991, Solow 1986). If the original data do not follow any simple distribution (multimodal Gaussian or lognormal) *disjunctive kriging* provides a nonlinear, distribution-dependent estimator (Armstrong and Matheron 1986*a*, 1986*b*, Olea 1991, Rivoirard 1994). When the data are not normally

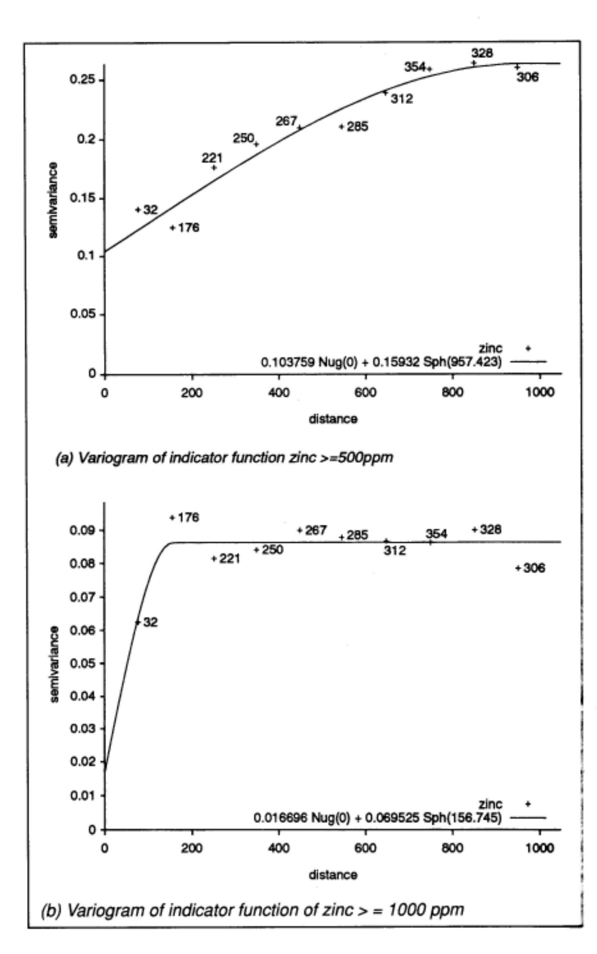


Figure 6.16. Indicator variograms of zinc: (*a*) for > 500 ppm level, (*b*) for > 1000 ppm level

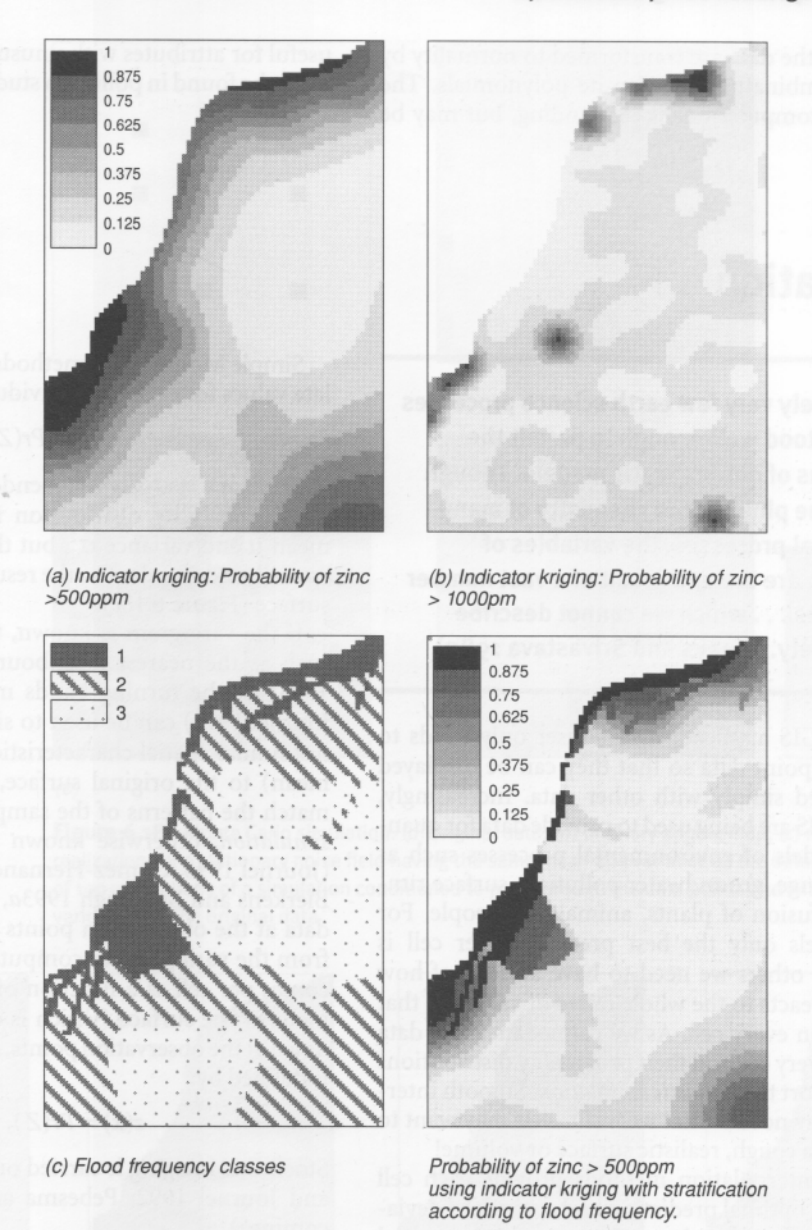


Figure 6.17. (*a*) Indicator kriging 500 ppm level; (*b*) idem 1000 ppm level; (*c*) flood frequency clasess; (*d*) indicator kriging with stratification in flood frequency classes 1 and 2 + 3 (500 ppm)

distributed the data are transformed to normality by a linear combination of Hermite polynomials. The method is computationally demanding, but may be useful for attributes with unusual distributions, such as can be found in pollution studies (e.g. see Burrough 1993).

Simulation

Unfortunately very few earth science processes are understood well enough to permit the applications of deterministic models. Though we know the physics and chemistry of many fundamental processes, the variables of interest . . . are the end result of a vast number of processes . . . which we cannot describe quantitatively. (Isaaks and Srivastava 1989)

For many GIS applications, the user only needs to interpolate point data so that they can be displayed or combined simply with other data. Increasingly, however, GIS are being used to provide data for quantitative models of environmental processes such as climate change, groundwater pollution, surface run-off, the diffusion of plants, animals, or people. For some models only the best prediction per cell is needed; for others we need to have an idea of how the model reacts to the whole range of variation that is possible in every cell. As we cannot measure data values for every cell and their probability distributions we must resort to *stochastic simulation*. Smooth interpolation may not be what we want—we may want to interpolate a rough, realistic surface or volume!

Kriging interpolation methods provide each cell with a local, optimal prediction and a standard deviation that depends on the variogram and the spatial configuration of the data. In order to examine how a given numerical model might react to input values larger or smaller than the local mean, we might compute upper and lower bounding surfaces from the the original interpolation and its standard deviations. These surfaces, however, only provide smooth estimates of possible values and for evaluating the sensitivity of models in which spatial contiguity is important (groundwater models, erosion, run-off, diffusion models) it is much better to have error data in which each cell has a unique and an equiprobable value.

Simple Monte Carlo methods can be used to simulate values for each cell individually:

$$z(\mathbf{x}) = Pr(Z) \qquad 6.29$$

where Z is a spatially independent, normally distributed probability distribution function (PDF) with mean μ and variance σ^2, but this treats every cell as spatially independent so the result is a stationary noise surface (Figure 6.18*c*).

If the variogram is known, then several methods such as the nearest neighbour method (Heuvelink 1993) or the turning bands method (Deutsch and Journel 1992) can be used to simulate a surface that has similar spatial characteristics (nugget, sill, range, mean) to the original surface, but which does not match the patterns of the sampled area. *Conditional simulation*, otherwise known as *stochastic imaging* (Journel 1996, Gomez-Hernandez and Journel 1992, Bierkens and Burrough 1993*a*, 1993*b*) combines the data at the observation points with the information from the variogram to compute the most likely outcomes per cell as a function of the variogram parameters. The surface (which is discontinuous, is tied down at the observation points, and varies in between) is defined by:

$$z(\mathbf{x}) = Pr(Z), \gamma(\mathbf{h}) \qquad 6.30$$

Stochastic imaging is carried out as follows (Deutsch and Journel 1992; Pebesma and Wesseling (forthcoming)).

1. Set up the usual equations for simple kriging with an overall mean.
2. Select an unsampled data point at random. Compute kriging prediction and standard deviation using data from the data set in the locality of the cell.
3. Draw a random value from the probability distribution defined by the prediction and standard deviation. Add this to the list of data points.
4. Repeat steps 2–3 until all cells have been visited and the simulation of one realization is complete.

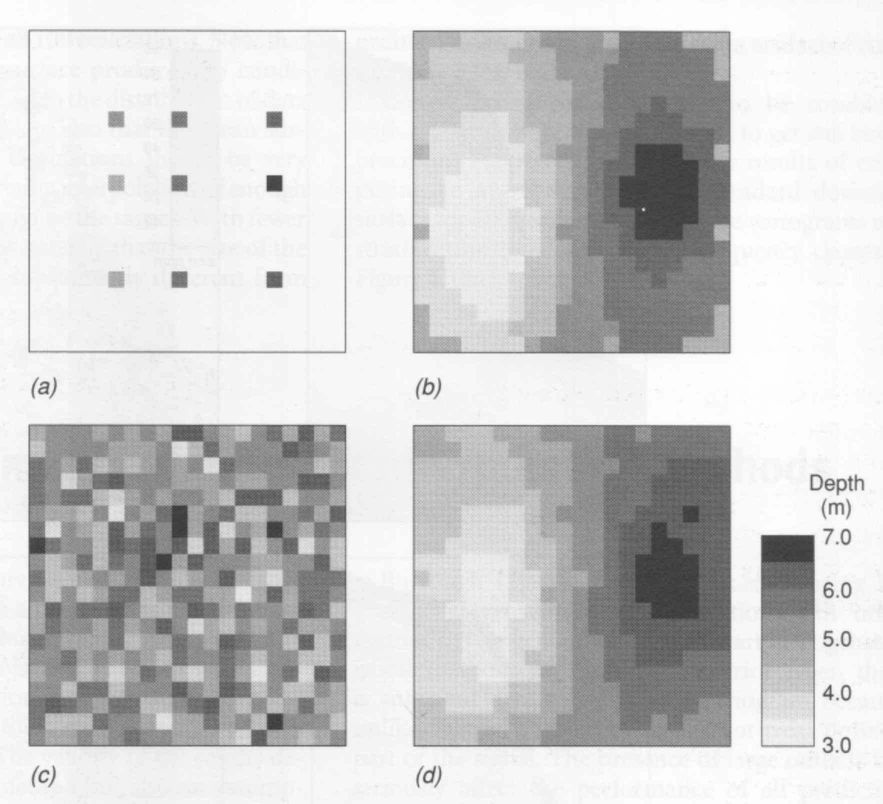


Figure 6.18. Monte Carlo simulation: (*a*) original data; (*b*) ordinary kriging map; (*c*) mean of 100 realizations of a stationary noise field having the same mean and variance as the data; (*d*) mean of 100 realizations of a simulation conditioned to pass through the data points and with the same variogram as the original data

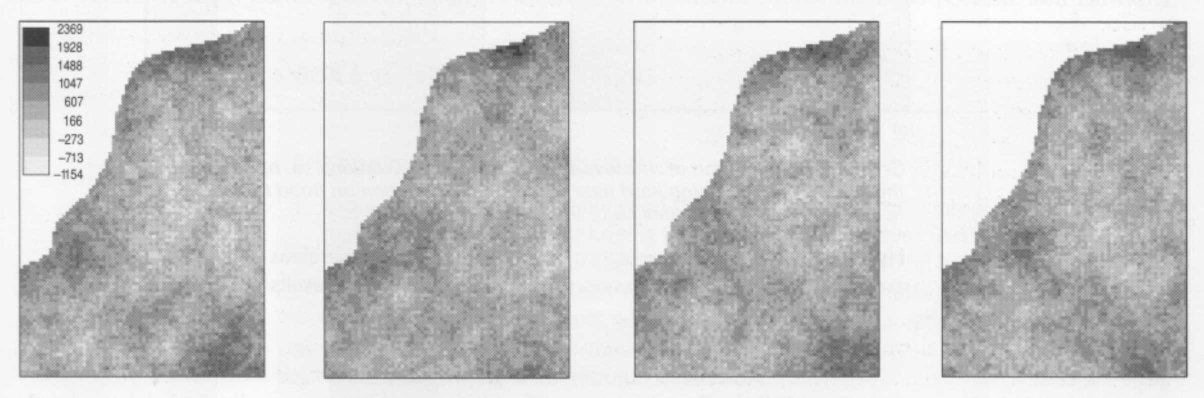


Figure 6.19. Four of 100 realizations of zinc levels produced by conditional simulation

5. Repeat steps 1–4 until sufficient realizations have been created.
6. Compute surfaces for the mean value and the standard deviation from all realizations if required.
7. Run the environmental model with each realization to see how results vary with the different inputs.

Conditional simulation is more computer intensive than either kriging or co-kriging. For example, ordin-

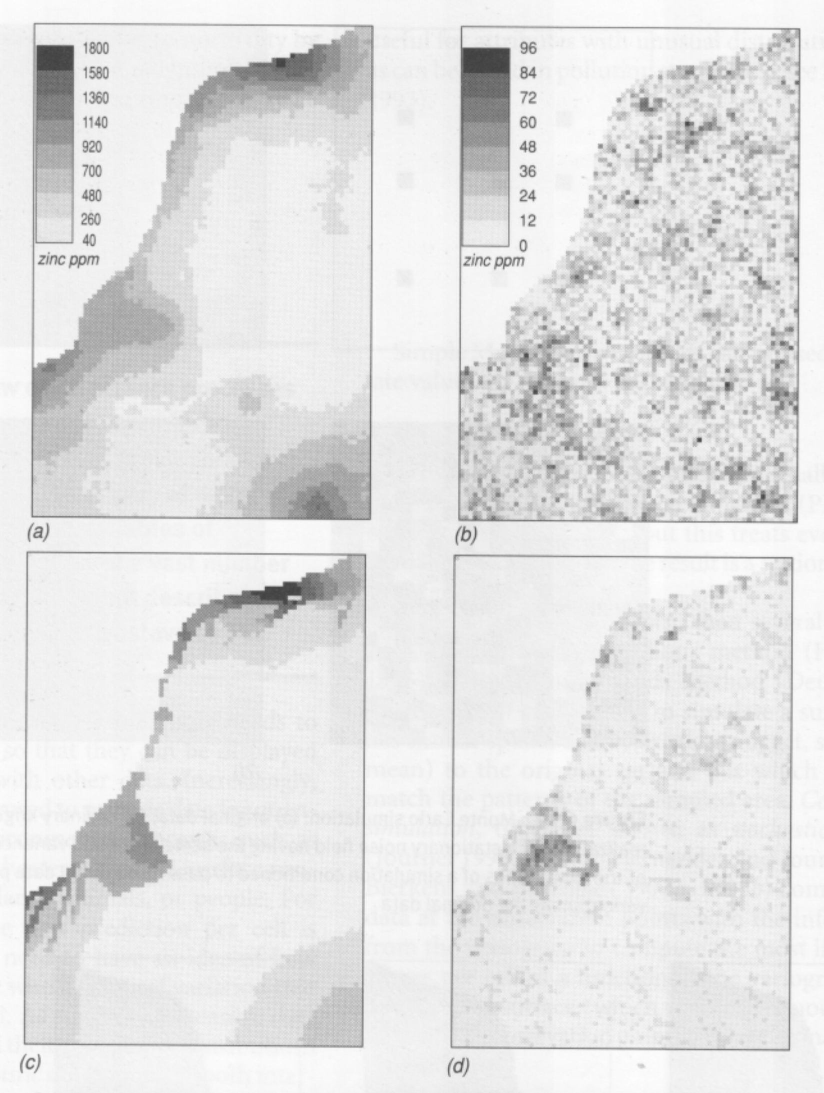


Figure 6.20. Conditional simulation can be carried out for whole areas or can be stratified according to flood frequency classes, which yields much preciser results

ary point kriging (Figure 6.6) on a 100 Mhz Pentium computer with 16 Mbyte RAM takes about 1 minute but conditional simulation takes about 2.2 minutes *per realization* on the same machine, or 3.5 hours in all for 100 realizations. It is sensible to use the method when the variogram is known and when most likely estimates, rather than smoothed averages, are needed to fill data cells for modelling.

The differences between ordinary kriging, simple simulation, and conditional simulation are shown in Figure 6.18. Usually one simulates at least 500 *realizations* of the random surfaces, which can be used as different inputs for a Monte Carlo analysis of models (see Chapter 10). Figure 6.19 shows four realizations from only 100 simulations of the variation of zinc content and Figure 6.20 shows the mean and standard

deviation surfaces for of all 100 realizations. Note that the standard deviation surface produced by conditional simulation is not tied to the distribution of data points, as with kriging. Note also that the mean surface of the conditional simulations should be very similar to the ordinary kriging interpolation (if enough replicates are used it should be the same). With fewer than 500 realizations it is possible that the size of the standard deviations are substantially different from ordinary kriging, and this could be an artifact of computing too few realizations.

Conditional simulation can also be combined with soft information in an attempt to get still better precision. Figure 6.20*c*,*d* shows the results of computing an average surface and standard deviation surface for 100 realizations using the variograms and stratification for the two flood frequency classes of Figure 6.14*a*.

The relative merits of different interpolation methods

Geostatistical methods are not one, but a wide range of techniques that rely on an understanding of the underlying spatial correlation structure of the data and use that to guide interpolation. Theoretical assumptions include ideas of stationarity, the intrinsic hypothesis and normality, and these are sometimes difficult to meet with real data. The validity of the results depends not only on the method but also on assumptions about uniformity of your area. Geostatistical methods have the advantage of providing estimates for points or blocks; information on spatial anisotropy, nested scales of variation, and external information can be combined to get the most out of expensive data.

Burrough (1993*b*) reviews work comparing the quality of geostatistical interpolation with other methods. The general conclusions are that geostatistical methods are generally superior when there is sufficient data to estimate a variogram because, unlike splines, such methods do not treat noise as part of the signal. The presence of large outliers can seriously affect the performance of all prediction methods, largely by distorting the estimated variogram and several authors recommend the removal of outliers by methods of exploratory data analysis (EDA—Haslett *et al.* 1990, Gunnink and Burrough 1997) or the use of robust methods (Cressie and Hawkins

BOX 6.3. THE STEPS IN KRIGING

Kriging requires the following steps:

1. Examine the data for normality and spatial trends and carry out appropriate transformations. If using indicator kriging transform to binary values (0/1).
2. Compute the experimental variogram and fit a suitable model if the spatial variation is more autocorrelated than just pure nugget (white noise). If the data are not autocorrelated (100 per cent nugget variance) then interpolation is not sensible.
3. Check the model by cross-validation.
4. Choose for kriging or conditional simulation.
5. If kriging, use the variogram model to interpolate sites on a regular grid where the sites are either equal in size to the original samples (point kriging) or are larger blocks of land (block kriging). If conditional simulation, compute at least 100 realizations on the regular grid. From these compute average and standard deviation surfaces.
6. Display results as grid cell maps or by threading contours (not with conditional simulation), either singly or draped over other data layers (e.g. a DEM).
7. Input the results to the GIS and use them in conjunction with the other data.

Table 6.1. Summary of results of geostatistical interpolation

Method	Minimum value (ppm)	Maximum value (ppm)	Per cent area >500 ppm	Per cent area >1000 ppm	Per cent area >1500 ppm
Ordinary point kriging with isotropic variogram	119	1661	29.28	4.91	0.36
The same, but with transformation to natural logarithms	122	1348	20.24	1.45	0.00
Ordinary point kriging with anisotropic variogram	127	1202	31.43	3.92	0.00
Ordinary point kriging within major flood categories	116	1817	15.56	3.53	0.85
Co-kriging on ln(distance to river)	144	1300	16.20	0.97	0.00
Universal point kriging with a regression model on ln(distance) and elevation	114	1483	18.89	2.33	0.00
Conditional simulation with one general variogram	140	1606	28.89	3.96	0.21
Conditional simulation with stratification according to flood frequency and 2 variograms	105	1800	15.11	3.53	0.86

Table 6.2. Comparison of prediction standard deviations for different methods

Method	Minimum standard deviation	Maximum standard deviation
Flood frequency map class 1	423	423
class 2	177	177
class 3	105	105
Ordinary point kriging with isotropic variogram for whole area	119	329
Ordinary point kriging with anisotropic variogram for whole area	238	392
Ordinary point kriging within major flood categories (2 variograms)	62	354

1980). Ordinary kriging is least successful when abrupt boundaries are present, though stratification into clearly different areas may bring considerable improvement as shown in this chapter. Conventional choropleth mapping is generally poor at predicting site-specific values when spatial variation is gradual and splines can behave unpredictably. When data are sparse, however, the classification approach usually yields the best prediction (see Chapter 10).

Comparing the results of the different interpolation methods in Chapters 5 and 6

In Table 5.5 we compared the results of the deterministic methods of interpolation, and Table 6.1 does the same for the geostatistical methods. Although there are no absolute standards for comparison, such as an independent data set for validation, we note that both deterministic methods and geostatistical methods show a wide range in results (cf. Englund 1990). Gen-

erally all methods (deterministic and geostatistical) that use only the data from the point observations estimate larger proportions of the area above the critical threshold values given in the table than those that use external information such as flooding classes and distance/elevation regression models.

Those methods where a kriging standard deviation map can be easily computed seem to show that stratified kriging has produced by far the best results (Tables 6.1 and 6.2). In general, the greater the information from data points and external sources (flood frequency classes), the lower the prediction standard deviations. Geostatistical prediction is effective at reducing interpolation errors where spatial variation is continuous; in combination with the stratification, even better prediction errors can be achieved. Whether these results are generally true requires testing with an independent data set, and this is deferred until Chapter 10. The evidence strongly suggests that using geostatistics and soft data together greatly improves the predictive power of GIS.

Table 6.3 gives a comparative overview of all methods of interpolation discussed in both Chapters 5 and 6.

Using variograms to optimize sampling

This chapter has been about methods of interpolation, their advantages and disadvantages and their strengths and weaknesses. Clearly, even with the best methods one is severely limited by the data, and though many GIS users may not collect their own data, for those that do it is worth considering if different numbers of data or arrangements of data points might not yield better information. This idea is further explored in Chapter 10; here we show how the variogram can be used to relate sample spacing (and layout) to the size of the cells used for block kriging.

Note that in ordinary block kriging (equations 6.17, 6.18) the prediction errors of z_B are controlled only by the variogram and the sampling configuration. Therefore, once the variogram is known it is possible to design sampling strategies that will result in any required minimum interpolation error. In particular, the prediction error σ_B^2 for a block of land B depends on:

(i) The form of the variogram (linear, spherical, or other function), the presence of anisotropy or non-normality, and the amount of nugget variance or noise.

(ii) The number of neighbouring data points that are used to compute the point or block estimate.

(iii) Sampling configuration—which is most efficient—irregular sampling, or a regular square grid or a triangular grid?

(iv) The size of the block of soil for which the estimate is made—is it an area equivalent to the support or a larger block of land?

(v) How the sample points are arranged with respect to the block.

These points are reviewed by Burrough (1991*a*) based on original work by Webster and colleagues (reported in Webster and Oliver 1990); here we only consider points (ii) and (v), namely the number of neighbouring data points on a regular grid and their sample spacing needed to yield a maximum value of σ_e^2 for a given block.

Once the variogram is known, one can calculate the combination of block size and sample spacing on a regular triangular or rectangular grid to yield predictions of block averages with a given prediction error. The prediction variance of a block B is the expected squared difference between the kriging prediction of the block mean and the true value. The standard error of the block mean is the square root of the prediction variance of the block B. The prediction variance is given by:

$$\sigma_B^2 = E\{[z_B - \hat{z}_B]^2\}$$

$$= 2\sum_{i=1}^{n}\lambda_i\gamma(x_i,B) - \sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j\gamma(x_i,x_j) - \gamma(B,B). \qquad 6.31$$

Table 6.3. A comparison of methods of interpolation

Method	Deterministic/ stochastic	Local/Global	Transitions abrupt/ gradual	Exact Interpolator	Limitations of the procedure	Best for	Computing load	Output data structure	Assumptions of interpolation model
Classification	Deterministic 'soft' information	Global	Abrupt if used alone	No	Delineation of areas and classes may be subjective. Error assessment limited to within-class standard deviations	Quick assessments when data are sparse Removing systematic differences before continuous interpolation from data points	Small	Classified polygons	Homogeneity within boundaries
Trend surfaces	Essentially deterministic (empirical)	Global	Gradual	No	Physical meaning of trend may be unclear. Outliers and edge effects may distort surface. Error assessment limited to goodness of fit	Quick assessment and removal of spatial trends	Small	Continuous, gridded surface	Phenomenological explanation of trend, normally distributed data
Regression models	Essentially deterministic (empirical-statistical)	Global with local refinements	Gradual if inputs have gradual variation	No	Result depends on the fit of the regression model and the quality and detail of the input data surfaces. Error assessment possible if input errors are known	Simple numerical modelling of expensive data when better methods are not available or budgets are limited	Small	Polygons or continuous, gridded surface	Phenomenological explanation of regression model
Thiessen polygons (proximal mapping)	Deterministic	Local	Abrupt	Yes	No error assessment, only one data point per polygon. Tessellation pattern depends on distribution of data.	Nominal data from point observations	Small	Polygons or gridded surface	Best local predictor is nearest data point
Pycnophylactic interpolation	Deterministic	Local	Gradual	No, but conserves volumes	Data inputs are counts or densities	Transforming step-wise patterns of population counts to continuous surfaces	Small-moderate	Gridded surface or contours	Continuous, smooth variation is better than ad hoc areas

Linear interpolation	Deterministic	Local	Gradual	Yes	No error assessment	Interpolating from point data when data densities are high as in converting gridded data from one projection to another	Small	Gridded surface	Data densities are so large that linear approximation is no problem
Moving averages and inverse distance weighting	Deterministic	Local	Gradual	Not with regular smoothing window, but can be forced	No error assessment. Results depend on size of search window and choice of weighting parameter. Poor choice of window can give artifacts when used with high data densities such as digitized contours	Quick interpolation from sparse data on regular grid or irregularly spaced samples.	Small	Gridded surface, contour lines	Underlying surface is smooth
Thin plate splines	Deterministic with local stochastic component	Local	Gradual	Yes, within smoothing limits	Goodness of fit possible, but within the assumptions that the fitted surface is perfectly smooth.	Quick interpolation (univariate or multivariate) of digital elevation data and related attributes to create DEMs from moderately detailed data.	Small	Gridded surface, contour lines	Underlying surface is smooth everywhere
Kriging	Stochastic	Local with global variograms Local with local variograms when stratified. Local with global trends	Gradual	Yes	Error assessment depends on variogram and distribution of data points and size of interpolated blocks. Requires care when modelling spatial correlation structures.	When data are sufficient to compute variograms, kriging provides a good interpolator for sparse data. Binary and nominal data can be interpolated with Indicator kriging. Soft information can also be incorporated as trends or stratification. Multivariate data can be interpolated with co-kriging	Moderate	Gridded surface	Interpolated surface is smooth. Statistical stationarity and the intrinsic hypothesis.
Conditional simulation	Stochastic	Local with global variograms Local with local variograms when stratified. Local with global trends	Irregular	No	Understanding of underlying stochastic process and models is necessary	Provides an excellent estimate of the range of possible values of an attribute at unsampled locations that are necessary for Monte Carlo analysis of numerical models, also for error assessments that do not depend on distribution of the data but on local values.	Moderate-Heavy	Gridded surfaces	Statistical stationarity and the intrinsic hypothesis

where $\gamma(x_i,x_j)$ is the semivariance of the attribute between points x_i, x_j, taking account of the distance $\mathbf{x}_i - \mathbf{x}_j$ between them (and the angle in cases of anisotropy), $\gamma(x_i,B)$ is the average semivariance between x_i and all points within the block, and $\gamma(B,B)$ is the average semivariance within the block (i.e. the within-block variance).

Equation 6.31 shows that the prediction variances are not constant, but depend on the size of block B, the form of the variogram, and the distance between the data points (i.e. the configuration of sampling points in relation to the block to be estimated). Note that these variances do not depend on the observed values themselves (except through the variogram). We can compute values of σ_B^2 for different block sizes and square grid spacing and plot them as a map: see Figure 6.21. These curves can help us decide what intensity of sampling and sample spacing is needed (and how much the sampling campaign will cost) to ensure that σ_B^2 is less than a given maximum value. This could be useful when trying to link data from point observations to data collected as grid cell averages (e.g. remotely sensed data) or where point data are to be used to drive a quantitative process model that uses a predetermined spatial resolution.

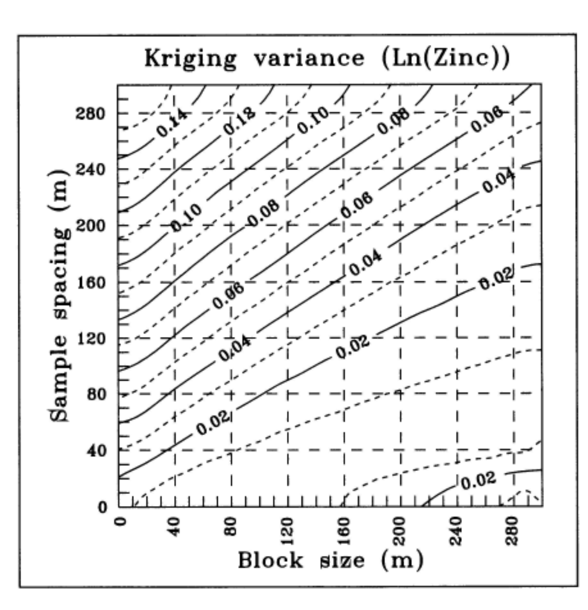


Figure 6.21. Isoline plot of equal prediction variances for ln(zinc) for different combinations of sample spacing and block size

Optimizing sampling when the variogram is unknown

If the variogram is not known, it can be estimated approximately by a reconnaissance method known as *nested sampling* (Webster and Oliver 1990). Clusters of samples are laid out over the area of interest in such a way that pairs of samples are located at short distances from each other, these pairs are at a larger, known distance apart, and these groups are still further apart, and so on. Nested analysis of variance is used to estimate how the cumulative variance varies with mean sample spacing, which was shown by Miesch (1975) to be equivalent to the variogram. Nested sampling is insufficient to provide sound estimates of the variogram for interpolation, but experience shows that the results of nested sampling can indicate near-optimal sample spacing for regular mapping by interpolation. Consequently, when variograms are unknown it can be cost-effective first to perform a nested sampling to determine the best sample spacing before regular sampling. Nested sampling can be made more efficient by the unbalanced approach which omits some replicates at the closer sample spacings (Webster and Oliver 1990).

Sources of software for geostatistical interpolation

All the interpolations and simulations presented in this chapter were carried out using GStat (Pebesma 1996) and PCRaster (Wesseling *et al.* 1996). Other useful sources of cheap or free geostatistical software such as GSLIB (Deutsch and Journel 1992) and others are given in Varekamp *et al.* (1996). Some statistical packages (e.g. SPLUS, GENSTAT, etc.) also include geostatistical methods. Pannatier (1996) has published an excellent set of interactive, Windows-based programs for modelling variograms and covariograms. The reader is advised to beware of incompletely documented software in general GIS packages!

Questions

1. Explain why it is so important in kriging to have a good model of the variogram.

2. Compare ordinary point kriging and thin plate splines as methods for interpolating elevation data to make a DEM.

3. Examine the costs and benefits of the different ways in which kriging interpolation can be assisted by using extra hard and soft information (hint—read Chapter 10).

4. Discuss ways of using indicator kriging to interpolate presence/absence data derived from social science surveys or vegetation studies.

5. Explain how you might use block kriging to ensure that point data are interpolated to a grid that has the same spacing and level of spatial generalization as a Thematic Mapper remote sensing image.

6. Explain to a decision-maker why it is worth having interpolated data with a known level of uncertainty.

Suggestions for further reading

BURROUGH, P. A. (1991). Sampling designs for quantifying map unit composition. In M. Mausbach and L. Wilding (eds.), *Spatial Variabilities of Soils and Landforms*, International Soil Science Society Working Group of Soil & Moisture Variability in Time and Space/American Society of Agronomy, the Crop Science Society of America, and the Soil Science Society of America, pp. 89–126.

—— (1993). Soil variability: a late 20th century view. *Soils & Fertilizers*, 56: 529–62.

CRESSIE, N. (1991). *Statistics for Spatial Data.* Wiley, New York.

DEUTSCH, C., and JOURNEL, A. G. (1992). GSLIB Geostatistical Handbook. Oxford University Press, New York.

ISAAKS, E. H., and SRIVASTAVA, R. M. (1989). *An Introduction to Applied Geostatistics.* Oxford University Press, New York.

PANNATIER, Y. (1996). *Variowin: Software for Spatial Data Analysis in 2D.* Statistics and Computing, Springer Verlag, Berlin.

VAREKAMP, C., SKIDMORE, A. K., and BURROUGH, P. A. (1996). Spatial interpolation using public domain software. *Photogrammetric Engineering and Remote Sensing*, 62: 845–54.

SEVEN

The Analysis of Discrete Entities in Space

The aim of GIS is not just to create a database of digital representations of geographical phenomena, but to provide means of selecting, retrieving, and analysing them. This chapter explains the methods available for dealing with crisp entities ('things')—how they can be selected from the database in terms of their attributes, how new attributes can be computed ('modelled') using the rules of Boolean logic and mathematics to yield useful groups or classes, or to generalize complex map images. Many of these procedures are not really spatial because they only affect the attributes and not the size, shape, or form of the spatial entities, which can be any geographical primitive—point, line, area, or pixel. Spatial analysis begins with the determination of spatial inclusion or exclusion, and with the intersection of lines and areas of different kinds to yield new entities. Spatial interactions are not just limited to the boundaries of existing entities but may be extended to include neighbourhood functions such as crow's flight distances, topological proximity, and distance over networks such as roads or rivers. These procedures are illustrated by examples from meteorology, archaeology, geodemographics, land evaluation, and planning.

Having gone to the trouble of collecting data and building a spatial database, the next issue is how to use these data to provide information to answer questions about the real world. This involves a wide range of methods of data manipulation from simple data retrieval and display to the creation and application of complex models for the analysis and comparison of different planning scenarios.

Some transformation capabilities, such as those necessary for data cleaning or updating, for changing scales or projections, or for interpolation were discussed in Chapters 4, 5, and 6. Geographical information systems, however, provide a large range of analysis capabilities that can be used in many ways. These analytical capabilities will usually be organized in modular commands so that each kind of analysis can be performed separately, or combined with others to build *data analysis models*. The actual user interface can be provided through typed commands, windows with preprogrammed buttons, or as statements in a

user-written higher programming language. The aim of this and the following chapter is not to instruct the reader on how to call up the program modules (which is system dependent) but to provide a proper understanding of the kinds of analytical functionality available and the tasks that can be accomplished.

The general problem of data analysis is stated in Figure 7.1. The user has a particular problem or query. The database contains information that can be used to answer the user's problem and will provide that answer in the form of a map, tables, or figures. To answer the query it is necessary to set up a formal set of data retrieval and analysis operations to recall the data, to compute new information and to display the results. This chapter and the following are about how these formal links can be defined and used.

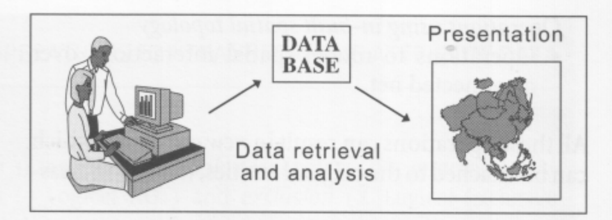


Figure 7.1. Data retrieval is the first step to visualizing information

Spatial analysis is more than asking questions

The kinds of analytical techniques that can be used on spatial data depend greatly on the data model and the representation that have been used. It is important to realize that different data models and different kinds of representation can require different approaches to the way spatial queries can be formulated. The fundamental question is whether the basic data model refers to *entities in space* or to the *continuous variation of an attribute over space*. In the case of entities, data retrieval and analysis concern the attributes, location, and connectivity of the entities and measures of the way they are distributed in space; in the case of continuous fields, data analysis concerns the spatial properties of the fields. The matter is made more complicated by the fact that continuous fields are usually *discretized* to a set of triangles or a regular grid, and the individual triangles or grid cells (or particular sets of contiguous triangles or grid cells) can also be treated as individual entities.

This chapter concentrates on the methods of data analysis (the links) that are most useful for dealing with *entities in space*, either in the relational of object-oriented model; the analysis of continuous fields is covered in Chapter 8. The fundamental axioms for data modelling and analysis were presented in Chapter 2. This and the following chapter demonstrate the applications of these axioms and how they can be translated into computer commands to solve practical problems.

The basic classes of operations for spatial analysis

In the entity model of objects in space three kinds of information are important, namely *what is it?*, *where is it?*, and *what is its relation to other entities?* The nature of an entity is given by its *attributes*, its where-abouts by its geographical *location* or *coordinates*, and the spatial relations between different entities in terms of *proximity* and *connectivity* (*topology*). The aspects of location, proximity, and topology distinguish geographical data from many other kinds of data that are routinely handled in information systems.

We distinguish the following basic classes of data analysis options for *entities*:

Attribute operations

- Operations on one or more attributes of an entity
- Operations on one or more attributes of multiple entities that overlap in space
- Operations on one or more attributes that are linked by directed pointers (object orientation)
- Operations on the attributes of entities that are contained by other entities (point in polygon)

Distance/location operations

- Operations to locate entities with respect to simple Euclidean distance or location criteria
- Operations to create buffer zones around an entity

Operations using in-built spatial topology

- Operations to model spatial interactions over a connected net.

All these operations can result in *new attributes*, which can be attached to the original entities, thereby increasing the size and value of the database. Certain operations also create new spatial entities, requiring the database to be expanded to include these new items. Data that have been retrieved from the database can be displayed on the screen, plotted as a paper map, or written as a computer file for future processing.

Operations on the attributes of geographic entities

As explained in Chapter 2, attributes are properties of entities that define what they are. They can be divided into three types—those that refer to location (the *geographical attributes* of latitude, longitude, and elevation) those that are simply attached as qualitative or quantitative descriptors of some non-spatial property, and those that are derived from the spatial properties of the entity itself. For example, the attributes of parcel number, name of the owner, and land cover describe non-spatial properties of a piece of land. The length of the fence bordering the road, the area, shape, and contiguity are attributes that are derived from the form of the piece of land.

As in conventional information systems, new attributes can be attached to entities as the result of a database operation. For example, a new attribute (or a new value of an attribute) can be computed for land parcels larger than a given size, or for those having owners that live abroad. For displaying information, the new attribute could be the colour or the symbol chosen to represent this kind of entity on the map. The new attribute can be derived by any legitimate method of logical and mathematical analysis, including operations on the proximity and topological properties of entities. Simple data retrieval can be seen as the creation of a new temporary attribute 'selected' when the set of attributes attached to an entity match the search criteria.

The process of selection or creation of new attributes can be formalized as follows. For any given location x, the value of a derived attribute U_i is given by:

$$U_i = f(A, B, C, D, \dots) \qquad 7.1$$

where $A, B, C, \dots$ are the values of the attributes used to estimate U_i. The function $f(\ )$ can be any of the following, singly or in combination:

(*a*) Logical (Boolean) operations
(*b*) Simple and complex arithmetical operations and numerical models
(*c*) Univariate statistical analysis
(*d*) Multivariate statistical methods or Bayesian statistics for classification and discrimination
(*e*) Multicriteria methods, AI-based methods: neural networks.

Logical (Boolean) operations on the attributes of one or more entities

When data have been encoded in a vector system using an overlay, feature plane, or layer structure then all data on that layer can be very simply retrieved by specifying the name of the layer. For example, if all roads are on layer 1, railways on layer 2, rivers on layer 3, and built up areas on layer 4, any single layer or combination can be easily retrieved and displayed.

Data retrieval using the attributes attached to individual entities

Entities can be selectively retrieved or reclassified on their attributes by using the standard rules of Boolean algebra which are incorporated in database languages such as SQL (see Date 1995, Chapter 6 for an introduction to data manipulation in relational database management systems). Boolean algebra uses the logical operators AND, OR, XOR, NOT to determine whether a particular condition is true or false (Box 7.1). Each attribute is thought of as defining a *set*. The operator *AND* (symbol $\wedge$) is the *intersection* of two sets—those entities that belong to both A and B; *OR* (symbol $\vee$) is the *union* of two sets—those entities that belong either to set A or to set B; *NOT* (symbol $\neg$) is the *difference* operator identifying those entities that

BOX 7.1. MATHEMATICAL OPERATIONS FOR TRANSFORMING ATTRIBUTE DATA

(*a*) *Logical operations.*
Truth or falsehood (0 or 1) resulting from *union* (∨ Logical OR), *intersection* (∧ Logical AND), *negation* (¬ Logical NOT) and *exclusion* (⊻ Logical Exclusive Or-XOR) of two or more sets.

(*b*) *Arithmetical operations.*
New attribute is the result of addition (+), subtraction (–), multiplication (*), division (/), raising to power (**), exponentiation (exp), logarithms (ln—natural, log—base 10), truncation, square root.

(*c*) *Trigonometric operations.*
New attribute is the sine (sin), cosine (cos), tangent (tan) or their inverse (arcsin, arccos, arctan), or is converted from degrees to radians or grad representation.

(*d*) *Data type operations.*
New attribute is original attribute expressed as a different data type (Boolean, nominal, ordinal, directional, integer, real, or topological data type).

(*e*) *Statistical operations.*
New attribute is the *mean, mode, median, standard deviation, variance, minimum, maximum, range, skewness, kurtosis*, etc. of a given attribute represented by *n* entities.

(*f*) *Multivariate operations*
New attribute is computed by a multivariate regression model.
New attribute is computed by a numerical model of a physical process.
New attribute is computed by a *Principal Component Analysis, Factor Analysis, Correspondence Analysis* transformation of multivariate data.
Entity is assigned to a given *class* (new attribute = class name) by methods of multivariate numerical taxonomy.
Entity is assigned a *probability* (based on statistical chance) by Discriminant Analysis, Maximum Likelihood or Bayesian techniques, of belonging to a given set.
Entity is assigned a *fuzzy membership value* for a given set.
Entity is assigned to a class using neural network methods.

belong to *A* but not to *B*, and *XOR* (symbol ⊻) is the *exclusive OR*, or the set of objects that belong to one set or another, but not to both. These simple set relations are often portrayed visually in the form of Venn diagrams (Figure 7.2). Note that logical operations can be applied to all data types, be they Boolean, nominal, ordinal, scalar, or directional.

Two simple examples illustrate the principles. Consider first a spatial database used by a real estate agent. A typical retrieval query from a prospective buyer might be the following: 'Please show me the locations of all houses costing between \$200 000 and \$300 000 with 4 bedrooms and plots measuring at least 300 m²'. If the data set contains the attributes 'cost', 'number of bedrooms', 'area of plot', and location, a map of the desired premises is easily produced by a multiple AND query on the specified attributes to highlight the matching plots:

```
IF COST GE $200 000 AND COST LT $300 000 AND
N BEDROOM = 4 AND PLOT AREA GE 300 THEN
ITEM = 1 ELSE ITEM = 0
```

The selected entities are given a Boolean value of 1 (true) if they match the specifications, and a 0 (false) if not. Display of the results follows by assigning a new colour to entities with ITEM = 1 (cf. Plate 2.1).

Now consider a query in land suitability classification. In a database of soil mapping units, each mapping unit may have attributes describing texture and pH of the topsoil. If set *A* is the set of mapping units

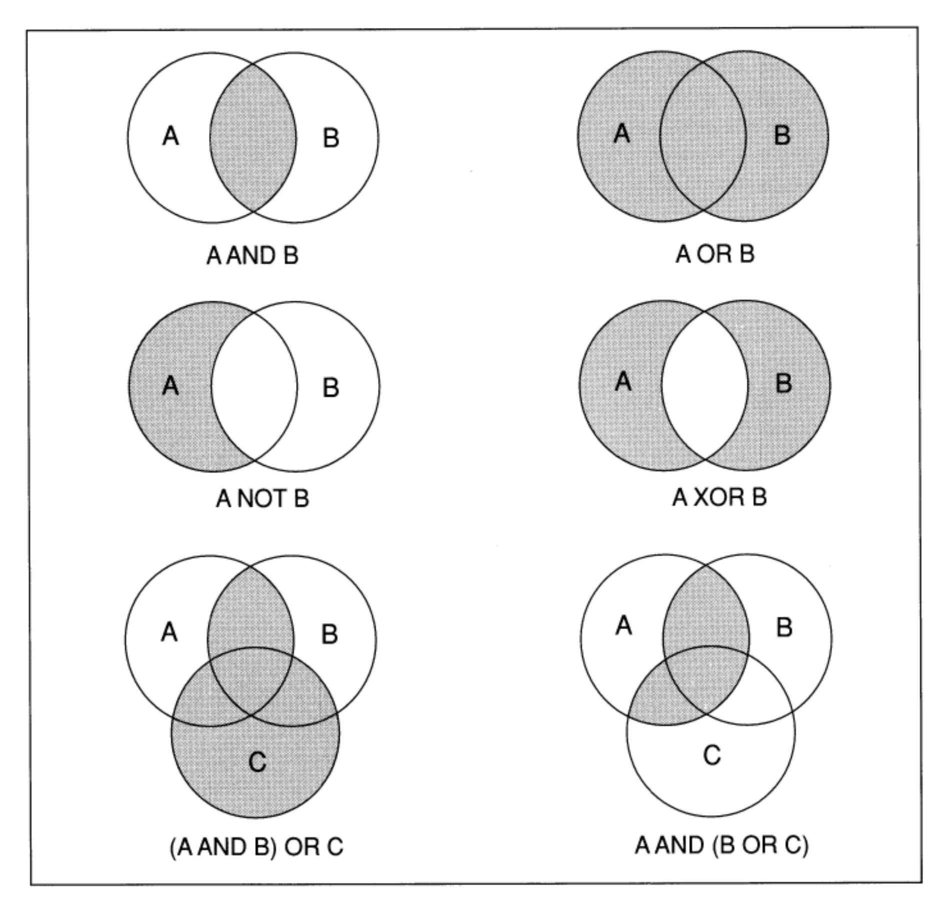


Figure 7.2. Venn diagrams showing the results of applying Boolean logic to the union and intersection of two or more sets. In each case the shaded zones are 'true'

called *Oregon loam* (nominal data type), and if set *B* is the set of mapping units for which the top soil pH equals or exceeds 7.0 (scalar data type), then the data retrieval statements work as follows:

> *X* = *A AND B* finds all occurrences of Oregon Loam with pH >= 7.0
> *X* = *A OR B* finds all occurrences of Oregon loam, and all mapping units with pH >= 7.0.
> *X* = *A XOR B* finds all mapping units that are either Oregon loam, or have a pH >= 7.0, but not in combination.
> *X* = *A NOT B* finds all mapping units that are Oregon loam where the pH is less than 7.0.

Selected entities can also be renamed and/or given a new display symbol (Figure 7.3*a*) by statements such as: 'Give the designation "Suitable" to all mapping units with soil texture = "loam" and pH >= 5.5'. This is a particular instance of the logical statement 'IF condition(C) THEN carry out specified task'.

Note that unlike arithmetic operations, Boolean operations are not commutative. The result of *A AND B OR C* depends on the priority of *AND* with respect to *OR*. Parentheses are usually used to indicate clearly the order of evaluation when there are more than two sets (Figure 7.2). For example if set *C* contains mapping units of poorly drained soil, then *X* = (*A AND B*) *OR C* returns all mapping units that are either (*a*) Oregon loam with a pH >= 7.0 or (*b*) units of poorly drained soil. The relation *X* = *A OR* (*B AND C*) returns (*a*) all Oregon loam mapping units and (*b*) those mapping units with a combination of pH >= 7.0 and poor drainage.

Note also that Boolean operations may require an exact match in attributes to return data and they take no account of errors or uncertainty unless that is specifically incorporated into the definitions of the sets.

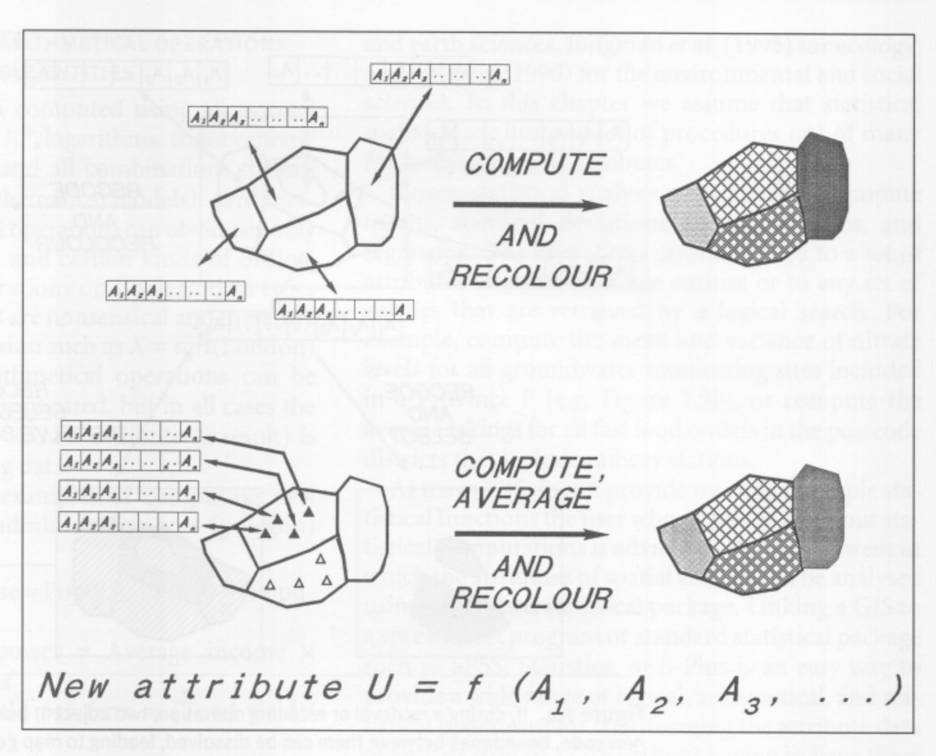


Figure 7.3. When retrieving entities on attributes alone, or computing new attributes from old, the spatial units of the map do not change shape, only colour or shading (top). The same occurs when point-in-polygon search is used to find enclosed point objects, that are used to compute area averages (bottom)

If the value of the attribute 'elevation' is set at 2000 m above sea level to define the class 'mountain' then hills with elevations up to 1999.999999999 . . . m will be rejected. This is not a problem with ordinal and nominal data types but it can present problems when working with scalar data types that represent quantities like elevation, pH, clay content, soil depth, atmospheric pressure, salinity, population, and so on, that are subject to various sources of measurement error and uncertainty. If the error bands on these data span the boundary values of sets then strict application of Boolean rules may yield results that are counter-intuitive (see Chapter 11).

Spatial aspects of Boolean retrieval on multiple attributes of single entities

Carrying out logical retrieval and reclassification on the non-spatial attributes of spatial entities has little effect on the map image, except in terms of symbolism and boundary removal. Computing a new attribute or condition requires the preparation of a legend and a recolouring or reshading of the selected entities (see Figure 7.3*a* and Plate 2.1). When selection leads to adjacent polygons receiving the same code it may be sensible to dissolve the boundaries between them, thereby achieving a form of map generalization (Figure 7.4). Figure 7.5 illustrates the use of this option to simplify a complex soil map.

Boolean operations are not only applicable to the non-spatial attributes of the geographical elements, because they can also be applied to the geographic location and attributes derived from the spatial or topological properties of the geographic entities. For example, one might wish to find all mapping units exceeding 5 ha in areas having soil with clay loam texture in combination with a pH > 7.0. More complicated searches may involve the shapes of areas, the properties of the boundaries of areas or the properties of neighbouring areas such as the areas of woodland bordering urban areas. In these cases, the results of the search would have an effect on the spatial patterns.

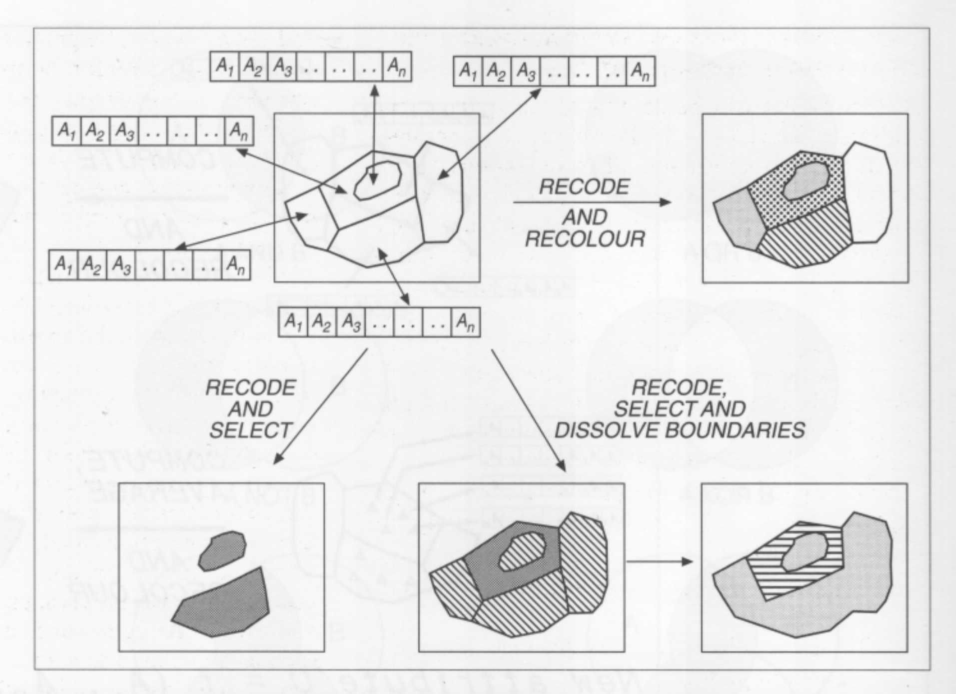


Figure 7.4. If, during a retrieval or recoding operation, two adjacent polygons receive the same new code, boundaries between them can be dissolved, leading to map generalization

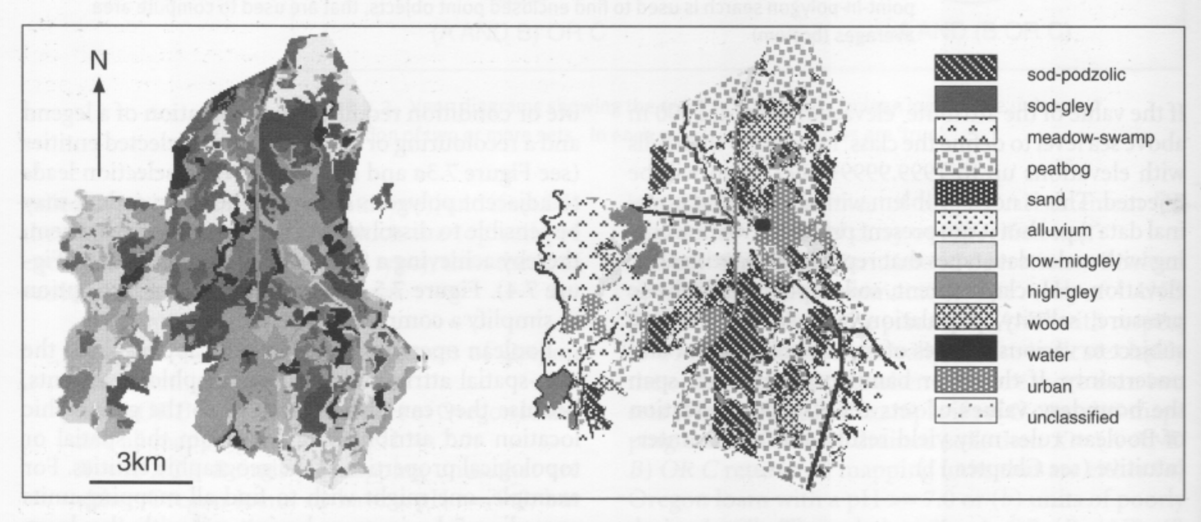


Figure 7.5. Using reclassification as a means of map generalization. Left: original soil map with 95 different units; right: reclassified map with 12 units. Note that reclassification preserves the original geometry and that the legend and shading codes only apply to the simplified map

Simple and complex arithmetical operations on attributes of single entities

New attributes can be computed using all normal arithmetical rules (+, –, /, *, logarithms, trigonometric functions, exponents, and all combinations thereof including complex mathematical models). Arithmetical and trigonometrical operations can obviously only be used on scalar data and certain kinds of ordinal data. Arithmetical operations on Boolean data types and nominal data types are nonsensical and therefore not allowed (an expression such as $X = sqrt(\text{London})$ has no meaning). Arithmetical operations can be very simple, or very complicated, but in all cases the operation is the same—a new attribute (a result) is computed from existing data.

Some hypothetical examples of computing new attributes for a given administrative area (polygon) or point location are:

- Population increase = Population 1990 – Population 1980
- Total spending power = Average income × number of persons
- Average wheat yield per farm = (Total yield)/(Number of farms)
- Predicted wheat yield = *f*(crop)
  where *f*(crop) is a complex mathematical model that computes wheat yield as a function of the soil, moisture, nutrients properties of a site (point entity)
- Class allocation = *Result*(multivariate classification)
  where (multivariate classification) might be any statistical or multicriteria analysis of the numerical attributes on the entity.

For a set of river catchments the proportion of precipitation discharging through the outlet can be computed by dividing the annual precipitation for each catchment by the cumulative annual river discharge measured at the outlet.

Arithmetical operations can be easily combined with logical operators:

IF (A + B)/(C) ≥ TEST VALUE THEN
CLASS = GOOD

The statistical analysis of attributes

There is no space in this book to explain the principles of univariate and multivariate statistical analysis but readers requiring details of these methods should consult a standard text such as Davis (1986) for geology and earth sciences, Jongman *et al.* (1995) for ecology, and Haining (1990) for the environmental and social sciences. In this chapter we assume that statistical methods are but one set of procedures out of many for computing new attributes.

Simple statistical analysis can be used to compute means, standard deviations and correlations, and regression. The operations can be applied to a set of attributes attached to single entities or to any set of entities that are retrieved by a logical search. For example, compute the mean and variance of nitrate levels for all groundwater monitoring sites included in a Province P (e.g. Figure 7.3*b*), or compute the average takings for all fast food outlets in the postcode districts that include railway stations.

As many GIS do not provide more than simple statistical functions the user who wishes to carry out statistical computations is advised to choose a system in which the attributes of spatial entities can be analysed using a standard statistical package. Linking a GIS to a spreadsheet program or standard statistical package such as SPSS, Statistica, or S-Plus is an easy way to provide a wide range of logical, arithmetical, and statistical operations for transforming the attribute data of entities held in a GIS without having to have these computational tools in the GIS. Both spreadsheets and statistical packages include useful graphics routines for plotting graphs, histograms, and other kinds of statistical charts. The use of hypertext facilities to link database, graphs, and map displays is providing powerful exploratory data analysis tools by which users can examine patterns in the probability distributions and correspondences in attribute data in relation to the spatial distribution of the entities to which they are attached (Gunnink and Burrough 1997, Haslett *et al.* 1990).

Most statistical packages provide at least the following procedures for statistical data analysis:

- Basic statistics—means, standard deviations, variances, skewness, kurtosis, maxima and minima, etc.
- Non-parametric statistics—median, mode, upper, and lower quartiles.
- Histograms, 2D and 3D scatter plots, Box and whisker plots, stem and leaf plots.
- Univariate and multivariate analysis of variance.
- Linear regression and correlation.
- Principal components and factor analysis.
- Cluster analysis.
- Canonical analysis.
- Discriminant analysis.

Numerical models

The range of arithmetical operations that can be applied to numerical attributes is unlimited. They are often used to compute the values of attributes that are difficult or impossible to measure, or which can be derived from cheap, readily available base data, such as data from censuses or natural resources surveys. Standard sets of mathematical operations that have been derived from empirical (regression) modelling are sometimes called *transfer functions* (Bouma and Bregt 1989) in disciplines like soil science and land evaluation. More complex sets of mathematical functions that represent a physical process such as crop growth, air quality, groundwater movement, pesticide leaching, epidemiological hazards, increase of population pressure, etc. are often referred to as *models*. Most GIS do not provide the functionality to program these complex models; instead they are used to assemble the data and export them to the model, which might reside on another computer on the network. Once the model results have been computed they are returned to the GIS as new attributes for display and evaluation. New developments, however, are providing GIS command interfaces with meta programming languages and specialized tools for spatial modelling (e.g. ARC-INFO—Batty and Xie 1994*a*, 1994*b*; PCRaster—Wesseling *et al.* 1996 or GRASS—Mitasova *et al.* 1996).

Neural networks, multicriteria evaluation, and fuzzy logic

All methods of deriving new attributes so far presented are *parametric*, which is to say that they assume that the definition of a new attribute can be expressed by a logical or numerical equation in which weights or parameters can be objectively assigned. Regression analysis and the calibration of numerical models are but two examples of ways in which the 'best' parameter values are chosen for classification or calculation. It is worth adding that in most cases the numerical models are also linear—i.e. there is a direct relation between a parameter value and its effect on the output.

These assumptions derive from classical, mechanical science. Parametric methods are difficult to use, however, in complex, non-linear situations where attribute values are not normally distributed and where causal or even statistical relations are tenuous. These difficult conditions surround many spatial data whose interrelations may violate many of the basic tenets of parametric methods and problems as simple as how best to classify complex spatial objects may be intractable. This problem has received particular attention in the classification of remotely sensed data into land cover classes (Lees 1996*a*).

Neural networks are providing new ways of classifying geographic entities (entities or pixels) into sensible groups (Fitzgerald and Lees 1996, Lees 1996*b*). In many cases the analyses are not really spatial, but require an entity to be assigned to a class on the basis of a non-statistical method of computation. A neural network is a processing device, implemented as an algorithm or computer program, that is said to mimic the human brain. It comprises many simple processing elements that are connected by unidirectional communication channels carrying numeric data. The elements operate only on local data but the network as a whole organizes itself according to perceived patterns in the input data. These patterns can be created by 'self-learning'—the system determines the 'best' set of classes for the data, or 'supervised classification' in which the system is supplied with a template of the required classification.

The application of neural networks to GIS is still in the research phase (see Lees 1996*b*), but indications are that these tools will provide useful and simple ways of dealing with complex data that otherwise might require extremely complicated modelling.

Multicriteria evaluation and fuzzy logic. Neural networks are not the only ways of dealing with complexity and non-linearity in spatial data. Methods of *multi-criteria* evaluation (Carver 1991) have been developed to provide a user with the means to determine new attributes that indicate alternative responses to problems involving multiple and conflicting criteria. In Chapter 11 we explore the particular use of fuzzy logic and continuous classification for spatial data analysis in GIS.

Examples of deriving new attributes for spatial entities

The results of computing new attributes or reclassification are usually displayed by reshading or recolouring the entities (Figure 7.2). As with Boolean selection, the spatial properties of the entities (location, shape, form, topology) do not change, except in the case that neighbouring entities are reclassified as being the same, when generalization can take place. Note that if the data are in raster form, these operations are carried out on each pixel separately, unless the data structure uses a map unit-based approach to raster data coding (see Chapter 3). The following sections give examples of spatial analysis based entirely on the derivation of new attributes.

Using empirical regression models

A simple example of a statistical model is provided by the linear regression of temperature in Swiss alps as a function of elevation. The relation between surface elevation and temperature in the Swiss Alps is almost linear and is of the form:

$$T = 5.697 - 0.00443 * E \qquad 7.2$$

where T is in degrees Celsius and E is elevation in metres. For an altitude matrix each grid cell is treated as a separate entity so applying equation (7.2) to all cells derives a gridded temperature map from the DEM. A similar example is given in Chapters 5 and 10 on the relations between pollution by heavy metals and independent attributes such as 'Distance from the river' or 'floodplain elevation' (equation 5.9).

Using multivariate clustering

Geodemographic segmentation (Webber 1997) is a method used by multinational marketing companies for classifying residential areas of Western countries into distinct neighbourhood types based on statistical information about the consumers who live in them. The spatial entities are provided by census districts, postal code districts, or mail order address units linked to the universe of national house addresses. This provides a fine spatial resolution of about seventeen addresses per spatial unit. Each spatial unit is characterized by four key criteria: age (young, middle, and old), income (high, middle, and low), urbanization (metropolitan, urban, suburban, rural), and family type (married couples with children, singles and childless couples, and pensioners). These attributes yield 108 possible combinations of classes which are recoded to ten core classes for the identification of characteristic socio-economic types. Multivariate methods of class allocation are used to assign a basic spatial unit to one of these ten classes which are also linked to empirical sales data and consumer preference attributes obtained by questionnaire. Simple logical retrieval of spatial units in terms of their class and attributes provides maps at local or national level that show the spatial distribution of market opportunities and brand preferences (Plates 2.5–2.8).

Using simple Boolean logic

In many parts of the world there are insufficient data to compute crop yields with numerical models of crop growth as a function of available moisture, energy, and water and so qualitative predictions based on simple rules may be the only useful way to assess land suitability for agricultural development. This is the philosophy behind the now classic FAO land evaluation procedure (FAO 1976, Beek 1978). Prescriptive land evaluation (Rossiter 1996, FAO 1976) is based on the simple idea that a landscape can be divided into basic entities called *mapping units*, separated by crisp boundaries. Soil survey is often the basis for this kind of landscape division, but alternative methods use landform, ecological zones, or vegetation communities. The idea is that once basic spatial entities have been mapped and defined in terms of representative attribute values their suitability for any given purpose may be determined by reclassification or by computing new attributes on the basis of existing information. The general procedure is called 'top-down' logic, because it starts with the presumption that global rules or physical models exist for translating primary units of information into units that can be used for a specific purpose. The procedure is exactly the same whether the data are stored and displayed as vector polygons or as grid cells; the only difference is the kind of spatial representation for the conceptual entity carrying the information.

An example of this rule-based reasoning is the following. In order to grow well, a crop needs a moist, well-drained, oxygenated fertile soil. Growing a monocrop leads to the soil being bare, or nearly bare for part of the year and in this time the soil should be able to

resist the effects of erosion by rain. The four attributes, available moisture, available oxygen, nutrients, and erosion susceptibility are known as *land qualities* (*LQi*), and they may be derived from primary soil and land data using simple logical transfer functions derived by agronomists and soil experts. The overall suitability of a site (its *land quality* for the use intended) is determined by the most limiting of the land characteristics—a case of worst takes all (FAO 1976, McRae and Burnham 1981).

Figure 7.6 illustrates the complete procedure using data from a soil series map and report and a digital elevation model of a small part of the Kisii District in Kenya (Wielemaker and Boxem 1982). The study area covers some 1406 ha (3750 × 3750 m) of the area mapped by the 1 : 12 500 detailed soil survey of the Marongo area (Boerma *et al.* 1974) and was chosen for its wide variety of parent material (seven major geological types), relief (altitude ranges from 4700 to 5300 feet a.s.l.—1420–1600 m) and soil (12 mapping units). Detailed soil survey information describes parent material, soil series, soil depth to weathered bedrock, stoniness and rockiness and tabular information relating soil series to land qualities. Each attribute was digitized as a separate polygon overlay and converted to a 60 × 60 array of 62.5 m square cells. The digital elevation model was obtained by interpolation from digitized contours and spot heights and converted to local relief (minimum 40 m, maximum 560 m). Information about the climate, the chemical status of the soil, the land use and cultural practices is also available (Wielemaker and Boxem 1982). Slope lengths (needed for estimating erosion) were interpreted from stereo aerial photographs and digitized as a separate overlay.

In this example we consider suitability for smallholder maize, which is determined by the land qualities *nutrient supply*, *oxygen supply*, *water supply*, and *erosion susceptibility*. These land qualities can be ranked by assigning values of 1, 2, 3 respectively to the following classes:

No limitation	assign 1
Moderate limitation	assign 2
Severe limitation	assign 3

The rules for deriving the land qualities are: water availability is a Boolean union (AND) of soil depth and soil series (*B*); oxygen availability and nutrient availability can be derived directly from the soil series information by recoding using a lookup table (*L*). Erosion susceptibility or hazard can be determined as a Boolean union (AND) of slope classes and soil series. Examples of the rules are:

If soil series is $\mathbf{S}_1$ *then assign nutrient quality* ***W3*** *from lookup table* $\mathbf{L}_w$.
If soil series is $\mathbf{S}_2$ *and slope class is 'flat' assign erosion susceptibility 1.*
If soil series is $\mathbf{S}_2$ *and slope class is 'steep' assign erosion hazard value 3.*

Once the individual land qualities have been assigned, the overall suitability per polygon or pixel is determined by the land quality with the most limiting (largest) value:

$$Suitability = maximum(LQ_{water}, LQ_{oxygen}, LQ_{nutrients}, LQ_{erosion})$$

This gives suitabilities of *poor*—at least one serious limitation; *moderate*—no severe but at least one moderate limitation and *good*—no limitations.

It is simple to repeat the analysis for the same area using other values of the conversion factors relating soil to the land qualities to see how the areas of suitable land may increase as limitations are dealt with by irrigation, mulching, or terracing. This *scenario exploration* can be achieved by replacing the real data with values of the land characteristics that represent a more degraded or an improved situation. In this way one can use the logical model to explore how the suitability of an area depends on the different factors. One must not forget, however, that the results are no better than the data and the insight in the land evaluation procedure allow.

Temporal Modelling of Attributes of Spatial Entities

A fundamental problem with simple rule-based modelling is that it takes no account of changes of the soil/landscape resource over time. We might expect that the growing of maize on bare plots might lead to soil erosion and degradation of the nutrient status, which even on initially suitable sites, might lead to reclassification to a poorer suitability class. So it would be useful to be able to estimate the suitability of the study area for smallholder maize for a time forty years hence, assuming that in the interim period the farmers have been growing maize over the whole of the area simply to feed the population which in 1985 was growing at more than 4 per cent per annum. Investigating which simple conservation methods the farmers can implement themselves in order to protect the soil from erosion, and determine in which parts of the area these

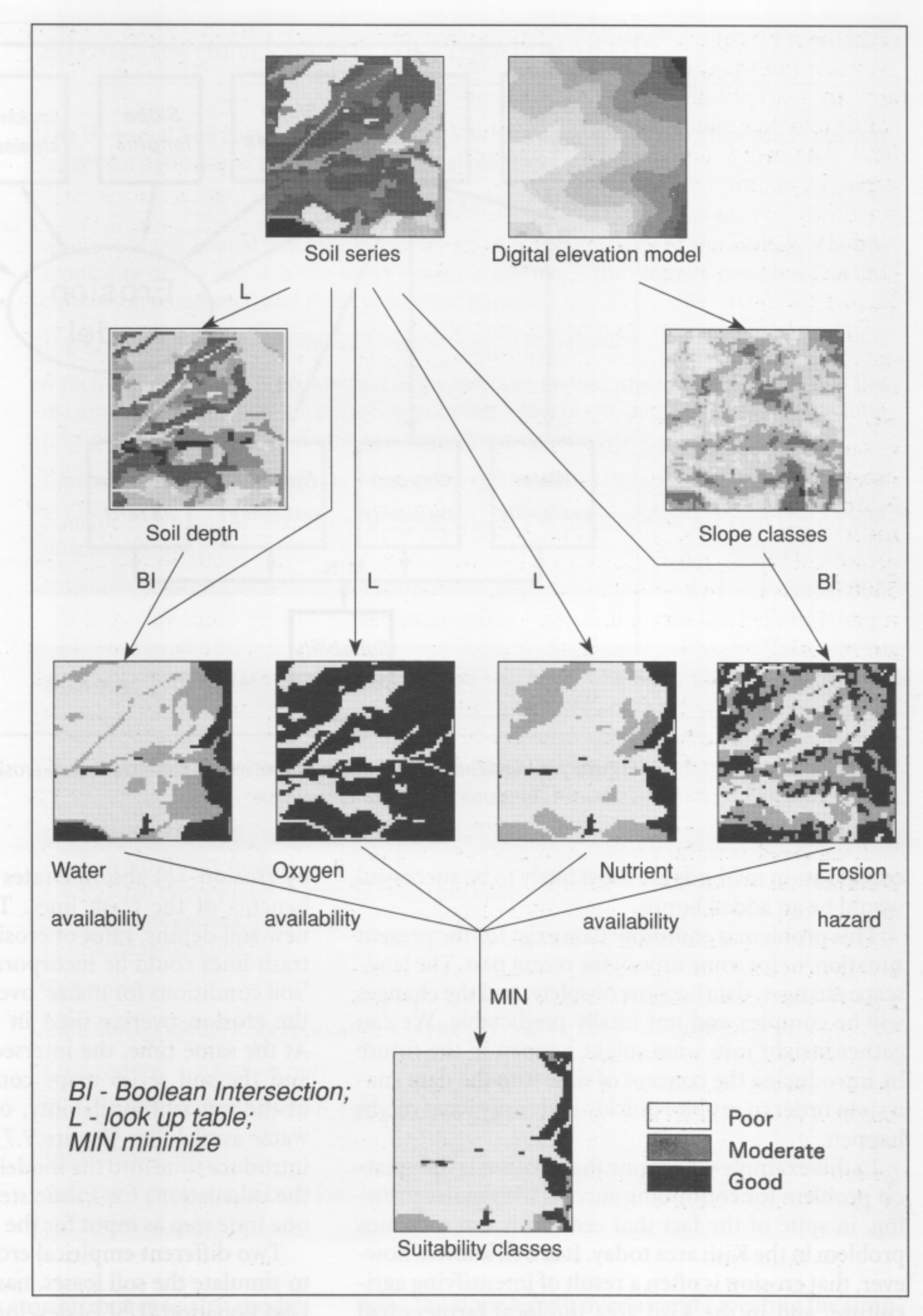


Figure 7.6. The flowchart of operations for 'top down' land evaluation for determining the suitability of land to grow maize

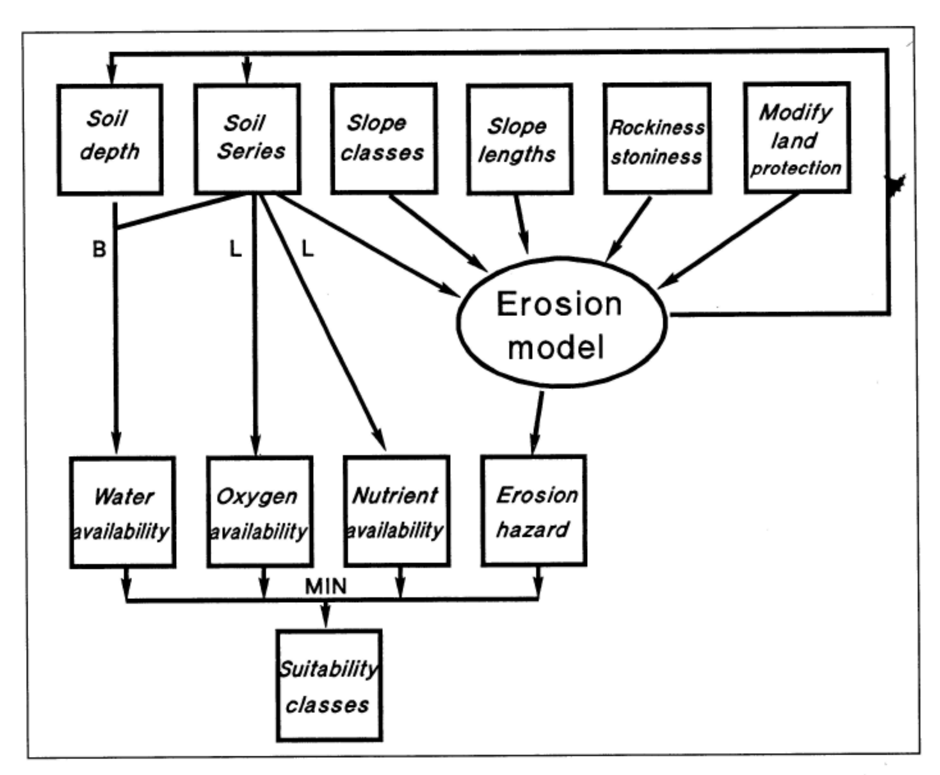


Figure 7.7. Flowchart of operations of land evaluation including an erosion model to update estimates of erosion susceptibility over time

conservation methods are most likely to be successful would be an added bonus.

This problem is common: data exist for the present situation, or for some time in the recent past. The landscape changes, data become obsolete, and the changes will be complex and not totally predictable. We can gather insight into what might happen in the future by introducing the concept of time into the data analysis in order to explore quickly and easily what might happen.

In this example we assume that erosion is the greatest problem for continuing success with maize growing, in spite of the fact that erosion is not a serious problem in the Kisii area today. It is well known, however, that erosion is often a result of intensifying agriculture, and in the Kisii area the local farmers had already been encouraged to make use of maize stalks to construct 'trash lines' along the contours in order to reduce soil losses (Wielemaker and Boxem 1982). So the main problem is one of how to estimate soil losses. This can be done by including an empirical erosion model into the land evaluation procedure in order to estimate (*a*) how soil depths might be reduced by erosion, (*b*) absolute rates of erosion, and (*c*) the benefits of the trash lines. The information about new soil depths, rates of erosion, and the benefits of trash lines could be incorporated into a new, future 'soil conditions for maize' overlay that would replace the erosion overlay used in the previous example. At the same time, the intersection of the new depth and the soil series maps could give new estimates of the nutrient availability, oxygen availability, and water availability (Figure 7.7). The simplest way to introduce time into the modelling process is to repeat the calculations for *t* time steps, using the results of one time step as input for the next.

Two different empirical erosion models were used to simulate the soil losses, namely the Universal Soil Loss Equation (USLE) developed by Wischmeier and Smith (1978) and the Soil Loss Estimation Model for Southern Africa (SLEMSA) developed by Stocking (1981). These empirical methods were used because they are well known, easy to compute, and the data for both was already in the database (see Box 7.2 and 7.3). The disadvantage of these empirical models is that they are a gross oversimplification of the real

BOX 7.2. THE UNIVERSAL SOIL LOSS EQUATION (USLE)

The Universal Soil Loss Equation (USLE—Wischmeier and Smith 1978) predicts erosion losses for agricultural land by the empirical relation:

$$A = R * K * L * S * C * P$$

where A is the annual soil loss in tonnes h^{-1}, R is the erosivity of the rainfall, K is the erodibility of the soil, L is the slope length in metres, S the slope in per cent, C is the cultivation parameter, and P the protection parameter.

The R, L, and S factors are derived from empirical regressions:

R factor. $R = 0.11\, abc + 66$ where a is the average annual precipitation in cm, b is the maximum day-precipitation occurring once in 2 years in cm, and c is the maximum total precipitation of a shower of one year occurring once in 2 years, also in cm.

L factor. $L = (l/22.1)^{1/2}$ where l is the slope length in metres (Wischmeier and Smith 1978)

S factor. $S = 0.0065s^2 + 0.0454s + 0.065$ where s is the slope as per cent (Smith and Wischmeier 1957).

BOX 7.3. THE SLEMSA EROSION MODEL

The Soil Loss Estimation Model for Southern Africa (SLEMSA—Elwell and Stocking 1982).

Control variables:

- E Seasonal rainfall energy (J/m²)
- F Soil erodibility (index)
- i Rainfall energy intercepted by crop (per cent)
- S Slope steepness (per cent)
- L Slope length (m)

Submodels

Bare soil condition $K = \exp[(0.4681 + 0.7663F).\ \ln E + 2.884 - 8.1209F]$

Crop canopy $C = \exp[-0.06i]$

Topography $X = L^{0.5}(0.76 + 0.53S + 0.076S^2)/25.65$

Output

Predicted mean annual soil loss (tonne ha^{-1}) $Z = K\,C\,X$

problems of estimating erosion, but specialist models are often location specific or require data that are not to be found in general purpose databases.

The erosion models are point operations; they compute the soil loss per entity (which in this case is a grid cell) without taking into account the soil losses and gains over neighbouring grid cells. The effects of the trash lines on the erosion were estimated by reclassifying the slope length overlay and repeating the simulations. In all, the simulations resulted in twelve new overlays, of new soil depth, of new rockiness and rate of erosion for both models with and without the effects of the trash lines. For a given model, subtraction of the overlay of the rate of erosion with trash lines from the overlay estimated using no trash lines gives a map of the benefit of terraces (Figure 7.8). Subtract-

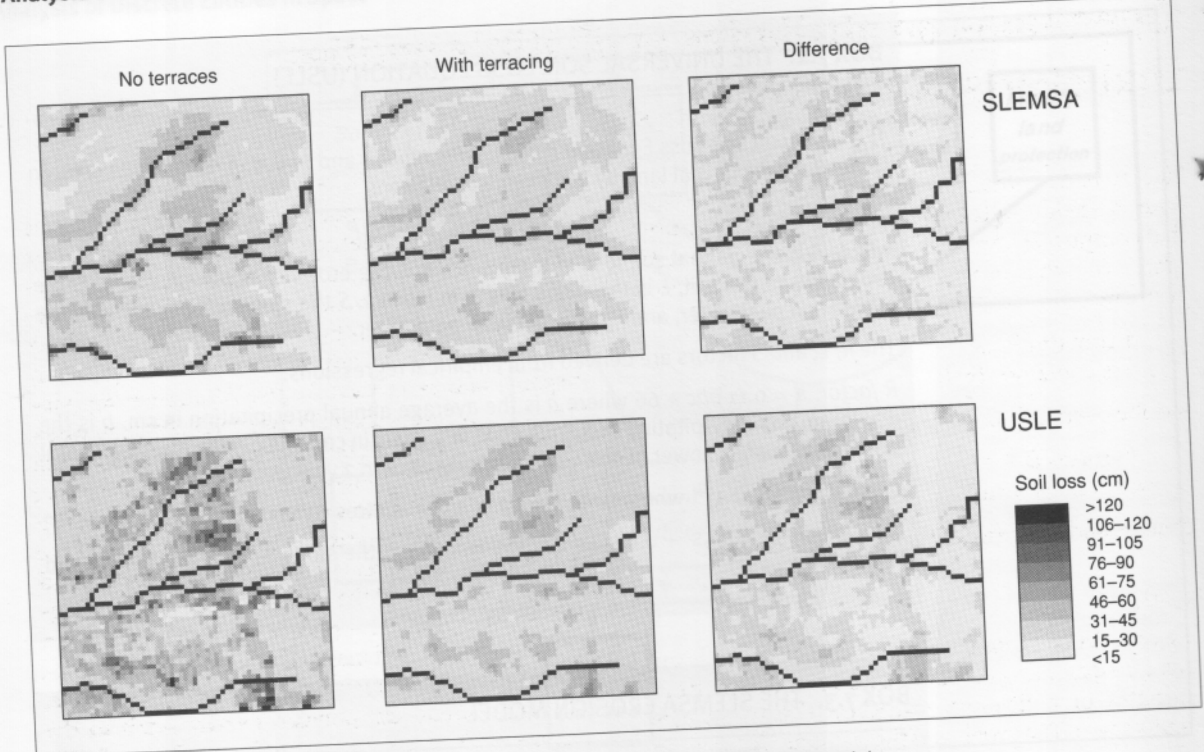


Figure 7.8. Results of evaluating the Flowchart of Figure 7.7 with two different erosion models (SLEMSA and USLE) and with different land protection practices (no protection and mulch terraces)

ing the overlays of the rates of erosion estimated under the same conditions by each of the two models yields an overlay of the difference in estimates provided by the models.

The results of the simulations suggest that there will be a slow degradation of the landscape in its suitability for maize over the period—the best areas decline from 14.1 per cent to 11.6 per cent, and marginal areas from 51.3 per cent to 50.6 per cent over the forty years. The simulations suggest that erosion will be concentrated in certain areas, and that in some of these the degree of erosion will be little modified by trash lines. The sensible conclusion would be to use these areas for some other, more permanent crop that would support the local population in other ways.

At first sight, the results seem quite convincing and realistic, but we have no way of knowing if they are anywhere near the truth. It is well known that empirical equations cannot easily be transported from one country to another, in spite of the many efforts that have been made to find truly 'universal' models. A further problem is that in all these overlay operations it has been assumed that the original data are absolute and invariate. This is clearly not the case, even for the relatively detailed soil survey in Kisii. The problem of error levels and error propagation in geographical information processing, particularly in the cartographic modelling methods using choropleth thematic maps is seldom discussed by the developers and users of many geographical information systems. The problem of errors is clearly so important with respect to the decisions that may be taken as a result of geographical analyses that this whole question is discussed more fully in Chapters 9 and 10.

Operations on attributes of multiple entities that overlap in space

Here we extend the discussion of operations on attributes to include attributes from two or more entities that completely or partially occupy or cover the same space. In other words we consider the inclusion problem:

A contains *B*, or
A is contained by *B*,

and the overlap and intersection problem:

A crosses *B*
A overlaps with *B*

where *A* and *B* are two different spatial entities.

Inclusion

The cases '*A* contains *B*' and '*A* is contained by *B*' are solved by extending the rules of Boolean algebra from attributes of entities to measures of how entities occupy space. The problem is the well-known 'point in polygon' issue, which was explained in Chapter 3. The first step in the analysis is to determine which entities are included or excluded in the location sense—e.g. 'Which restaurants are located in Soho?', 'Which groundwater observation wells have been drilled in formation X?' Once the entities have been selected and tagged, the procedures for attribute analysis can be applied, either per entity, or collectively. For example, the minimum and maximum water levels could be extracted for a given year for each groundwater well, or the average water level of all wells could be computed. The result of these computations can be used to tag the enclosing polygon, which can be displayed with a new colour, shading, or label (Figure 7.3*b*).

Other examples of applications of this kind of analysis are: from archaeology, 'determine the number of late-Iron Age burial sites in parish A', or 'retrieve all passage graves and determine the kinds of soil and landscape position where they occur', or from soil science: 'find all soil profiles located in unit *S1*, and compute the mean value of the clay content of the topsoil and its standard deviation'.

Entity overlap and intersection

In certain cases of '*A* contains *B* and *A* is contained by *B*' and with '*A* crosses *B* and *A* overlaps with *B*', where *A* and *B* are lines or polygons of different form, the first steps in logical retrieval define new areas or line segments. With polygons the process is known as *polygon overlay and intersection*, and it leads to the creation of new spatial entities. Figure 7.9 shows three different results, depending on whether the operation is spatial Boolean *Union* (Figure 7.9*a*), *covering* (Figure 7.9*b*), or *Intersection* (Figure 7.9*c*). Polygon overlay is used to answer questions such as 'Find the area of the City of Westminster that is covered by parks', in which the first overlay may show the boundaries of the administrative areas of urban London and the second is an overlay delineating different kinds of land cover. An example of the 'cookie cutter' overlay (Figure 7.9*c*) in a physical application is that if Map *A* is a map of soil types, and map *B* the boundary of a catchment, the result is a soil map of that catchment alone.

In some situations, polygon overlap leads to the creation of so-called *spurious polygons* (Figure 7.10). This is because of small errors in the digitizing of boundaries that are supposed to lie in the same place. The errors can result from placement errors of the digitizer puck, differences caused by paper stretch in the different source documents, or errors made during surveying. Paradoxically, attempting to digitize the lines more accurately with more data points does not solve the problem but actually makes it worse. There are several solutions. The first is to designate the boundaries on one feature layer as the dominant boundaries to which all others must defer. The second is to examine all the spurious polygons and eliminate all those have an area smaller than some critical threshold; here decisions must be made as to which of the larger polygons the area covered by the spurious polygons should be added. The third is to pass a smoothing window over all the coordinates of spurious polygons along the conjugate boundary zones and to compute a new, average boundary. This is frequently over-defined and can be simplified using the Douglas–Peuker algorithm (Douglas and Peuker 1973) or other means and smoothed for computing the boundaries of the new polygon entities and for display (Figure 7.10).

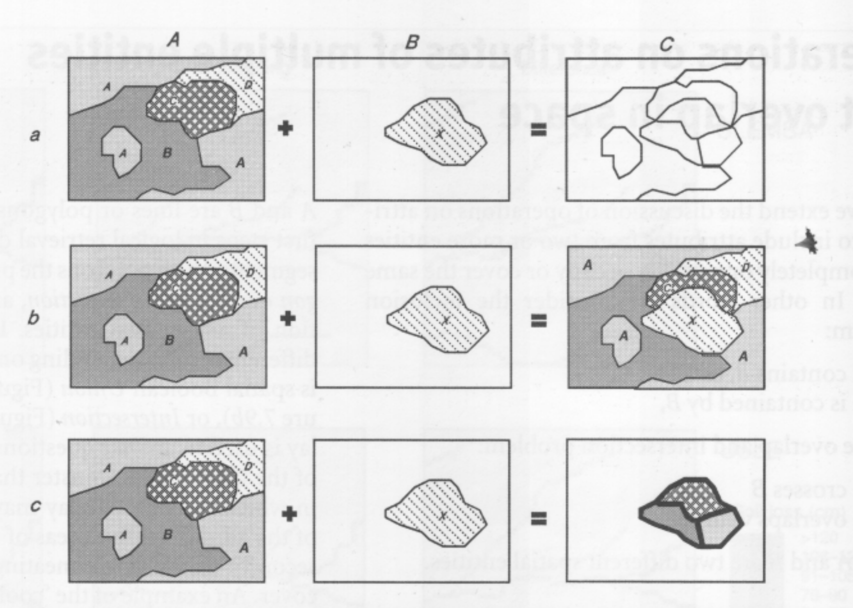


Figure 7.9. Polygon overlay leads to an increase in the number of entities in the database. (*a*) simple overlay—all boundaries are retained; (*b*) second map covers the first and changes the map detail locally; (*c*) the covering map is used to cut out a small part of the first map

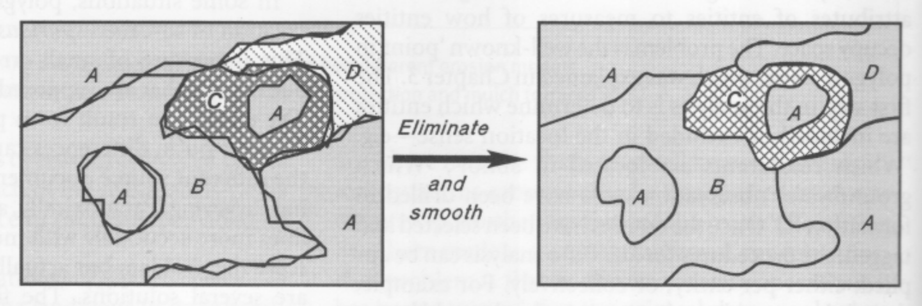


Figure 7.10. Polygon overlay can lead to a large number of spurious small polygons that have no real meaning and must be removed

Operations on one or more entities that are linked by directed pointers (object orientation)

Logical operations on lines and polygons that result in entities being split or removed from the database impose heavy computational costs on a spatial database because the number of entities may depend on the operations being carried out. In practical terms, if two simple polygons intersect to create a third, then the third and the attributes it inherits from the two original polygons must be added to the database. If reclassifying two adjacent polygons results in the removal of a common, unnecessary boundary, then two polygons must be removed and one added to the database. In practice, the number of polygons added or removed may be large and indeterminate, so it is difficult to say just how great this overhead is. To save modifying the original database, the changes are often only computed on a subset of the original data, and the results are stored in a separate file or folder.

In hybrid-relational GIS, adding and removing polygons means modifying both the spatial data and the attribute data separately. Modifying the spatial data is more than just adding or deleting an entry in a table because all the topological connections need to be recomputed. Most commercial GIS have adopted compromise solutions, taking care of the modification

of the spatial representations with their own software and using commercial relational database programs for storing the associated attributes.

The advantage of network-relational hybrid databases is that in principle there is no limit to the kind and number of analysis queries that can be defined. Object-oriented GIS attempt to get around these computational problems by incorporating a large amount of information to structure the data in such a way that data volumes do not greatly change as queries are carried out. This means that the most common data retrieval and analysis options need to be thought out beforehand, which is why constructing an object-oriented database can take much time.

Operations that depend on a simple distance between A and B (buffering)

Operations of the type '*A* is within/beyond distance *D* from *B*', where *D* is a simple crow's flight distance are carried out with the help of a *buffering* command. This is used to draw a zone around the initial entity where the boundaries of the zone or buffer are all distance *D* from the coordinates of the original entity. If that is a point entity, then the zone is a circle, if a straight line, a rectangle with rounded ends, an irregular line or polygon, and enlarged version of the same (Figure 7.11). The buffer is in effect a new polygon that may be used as a temporary aid to spatial query or that is itself added to the database. The determination of whether an entity is inside/outside or overlaps the buffer zone is then carried out using the operations just described (including the problems of polygon overlay!—see Figure 7.7) and logical or mathematical operations on those entities proceed as before.

Typical examples of using the zoning/buffering command with other analysis options are the following:

- 'Determine the number of fast food restaurants within 5 km of the White House.'
- 'Investigate the potential for water pollution in terms of the proximity of filling stations to natural waterways.'
- 'Compute the total value of the houses lying within 200 m of the proposed route for a new road.'
- 'Compute the proportion of the world population that lives within 100 km of the sea.'
- 'Compute the number of cattle grazing within 5 km of a waterhole.'
- 'Determine the potential amount of arable land within 1 hour's walk from a Neolithic village.'

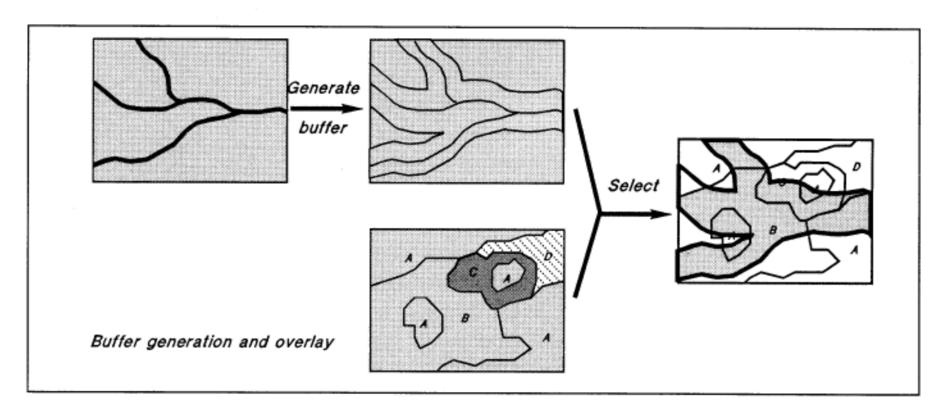


Figure 7.11. Generating buffer zones around exact entities such as points, lines, or polygons yields new polygons which can be used in polygon overlay to select defined areas of the map

Operations that depend on connectivity

These are operations in which the entities are directly linked in the database; the linkage can be spatial, as in the contiguity case where *A* is a direct neighbour of *B*, or the case where *A* is connected to *B* by a topological network that models roads or other lines of communication. Entities can also be linked by an internal topology, so that complex spatial entities are made up of sets of sub-entities, as is the case with object orientation.

The operations '*A* is a direct neighbour of *B*' and '*A* is connected to *B* by a topological network' are two versions of the same which, for topologically connected lines and polygons use explicit information from the spatial database (see Chapter 3) to determine how two entities or locations are connected. Inter-entity distances over the network or other measures of connectivity such as travel times, attractiveness of a route, etc. can be used to determine indices of interaction. These operations are much used for determining the location of emergency services or for optimizing delivery routes.

For example, when a boundary between two land cover polygons is also defined as a road, it is a simple matter to select those roads/boundary lines that have particular kinds of land cover on both sides. Such an analysis would easily distinguish rural roads (agriculture on both sides) from urban roads (built up areas on both sides) from coastal roads (sea on at least one side—to take account of sea dikes and breakwaters).

The analysis of connectivity over a topologically directed net is much used in automated route finding (car and truck navigation systems) and for the optimum location of emergency services. Attributes attached to the line elements representing roads, rivers, or rail links can identify the character of the connector. For example a road can be identified not only by its width, surface, class, and number of lanes, but also by its visual attractiveness or otherwise (potential tourist route) and traffic densities. Linking time series data on traffic densities over a day and a week to the route information provides a sound basis for estimating travel times for all hours of the day, factors that are important for the location of emergency services (Geertman and Ritsema Van Eck 1995). Figure 7.12 shows how the route from *A* to *B* over a network may depend on the attributes of the roads taken and may be quite different from that computed from a

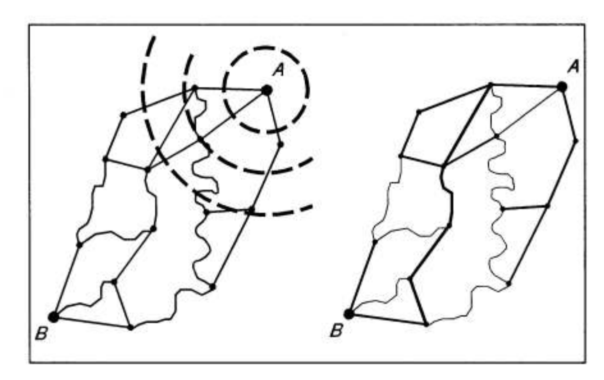


Figure 7.12. The analysis of transport times from *A* to *B* in terms of (*a*) crow's flight distance and (*b*) times along different routes in a network to determine expected travel times for different road conditions

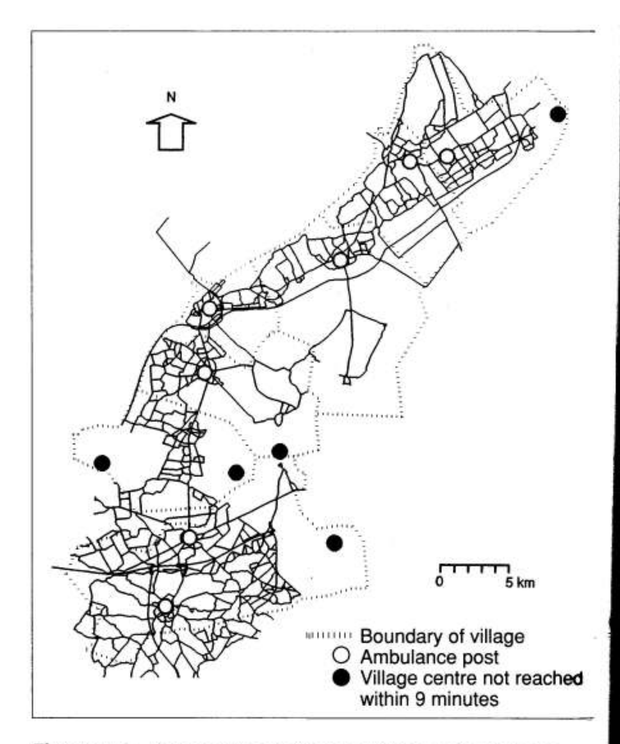


Figure 7.13. The results of a transport time analysis to see which suburban areas can be reached by ambulance within 9 minutes from the ambulance posts. Black lines show roads that can be reached within 9 minutes from the ambulance posts (open circles). Black circles show local centres that cannot be reached in that time. Pecked lines show outline of urban area

crow's flight path based on simple buffering. Plate 2.2 compares travel times computed from simple crow's flight distance obtained by buffering with those obtained from accumulated times over the connected road network: the comparison of travel times depends on the traffic regimes at different times of the day. This is part of analysis of the accessibility of built-up areas for emergency services (Figure 7.13—Ritsema Van Eck 1993). Plates 2.3 and 2.4 show results of an analysis of the travel times endured by commuters in the west of the Netherlands depending on whether they travel by private cars or public transport.

General aspects of data retrieval and modelling using entities

Spatial entities can be retrieved and new attributes can be computed by a wide range of logical and numerical methods. The numerical procedures can also be applied to inclusion and intersection problems and for proximity analysis and for analysis of relations over topological connections. The methods can be combined to create complex models for addressing many different kinds of spatial problem. Note that many data analysis operations are not commutative so the sequence in which the commands is executed is very important. While it can be very informative to sit in front of the computer browsing through a database to see what is there or how different procedures work (e.g. with Exploratory Data Analysis—Haslett *et al.* 1990, Gunnink and Burrough 1997) informal procedures are best for simple data retrieval and transformations. However, when a complex series of commands must be used frequently to retrieve and transform data it is sensible to create a structured command file (using a DOS or UNIX interface or a command language like the ARC-INFO AML) that can be reviewed, modified, and used by several persons. Such a set of commands constitutes a 'model' or a 'procedure' which can be stored in the GIS, referenced directly by an icon or a name, and used on other databases to carry out the same set of operations.

None of the methods of analysis presented in this chapter pay any attention to data quality or errors; there is a tacit assumption that all data and all relations are known exactly. In spite of this (which is a topic to which we return in Chapters 9 and 10), spatial modelling with GIS has great value for exploring different scenarios. Rather than letting land use planners loose in situations in which they have little or no experience (and the increasing pressure on land in developing countries is giving rise to situations in which few have the necessary expertise to handle) it is surely preferable to train them on digital models of the landscapes. Just as pilots learn to fly on flight simulators and architects and road designers build maquettes in order to broaden their experience and develop their ideas, so the land use planner can learn from 'mistakes' made on digital landscape models before irretrievable errors are made in the landscape itself.

Questions

1. Develop a simple entity-based model to analyse the effects of land use change annually.

2. Work out a GIS-based system for the optimum location of (*a*) fire stations, (*b*) banks, (*c*) health care services in cities.

3. Explore the advantages and shortcomings of using entity-based models for ecological modelling.

4. Develop a GIS system for helping to manage the demand for building materials required for constructing a new suburb.

5. Explore the advantages of entity-based GIS in (*a*) real estate management, (*b*) hydrological modelling, and (*c*) archaeological site investigations.

6. Design a GIS-based method for providing on-line information to tourists.

Further reading

BATTY, M., and XIE, Y. (1994*a*). Modelling inside GIS. Part 1: model structures, exploratory spatial data analysis, and aggregation. *International Journal of Geographical Information Systems*, 8: 291–307.

—— —— (1994*b*). Modelling inside GIS. Part 2: Selecting and calibrating urban models using ARC-INFO. *International Journal of Geographical Information Systems*, 8: 451–70.

GOODCHILD, M. F., PARKS, B. O., and STEYAERT, L. T. (1993). *Environmental Modeling with GIS*. Oxford University Press.

—— STEYAERT, L. T., PARKS, B. O., JOHNSTON, C., MAIDMENT, D., CRANE, M., and GLENDINNING, S. (eds.) (1996). *GIS and Environmental Modeling: Progress and Research Issues*. GIS World Books, Fort Collins, Colo., 486 pp.

EIGHT

Spatial Analysis Using Continuous Fields

The paradigm of continuous fields provides a rich foundation for spatial modelling, particularly when data are held in regular, square rasters (grids). Methods of Map Algebra allow mathematical operations to be carried out on whole raster overlays just as easily as if each overlay were only a single number, and this facilitates the writing of numerical models. Mathematical operations on continuous fields can be divided into point operations and spatial operations. Point operations are the same as those discussed for attributes in Chapter 7; spatial operations include spatial filtering, the computation of surface derivatives (slope, aspect, convexity), surface topology and drainage nets, spatial contiguity, linear and non-linear proximity determination, and properties of whole surfaces such as viewsheds, shaded relief, and irradiance calculations. This chapter explains each of these operations; the methods are illustrated by applications in hydrology, erosion, and surface runoff, optimizing timber extraction from a forest, and determining links between radiocaesium levels in organic soils (originating from the Chernobyl accident) as a function of proximity to flooding by major rivers.

As we explained in Chapter 2, there are two main ways of representing continuous fields. The first is the Delaunay triangulation (the TIN of digital elevation modelling); the second is the more common altitude matrix or grid used in raster GIS and image analysis. Delaunay networks are often used outside GIS to support the finite element modelling of dynamic flow processes in groundwater movement (MODFLOW—McDonald and Harbaugh 1988), discharge over floodplains (e.g. Gee *et al.* 1990) or air quality (Fedra 1996). Finite element modelling (FEM) is not usually part of the standard generic tool kit of most GIS, although Kuniansky and Lowther (1993) report the generation of finite element meshes by stand-alone macro programs in a vector GIS. Numerical models that use FEM are usually loosely coupled to the GIS, with the GIS being used to assemble the data and pass them to the model via an interface. The results from the model are returned to the GIS, converted to a square grid or contour lines for ease of handling, and then overlaid on digitized base maps for display.

Finite element modelling is outside the scope of this book. Here we describe the operations for spatial analysis of continuous fields that are represented by the regular square grid, where each attribute is represented by a separate overlay, and each grid cell is allowed to take a different, scalar value (Chapter 3). Note that although in many examples we often refer to an *altitude matrix* (i.e. a gridded digital elevation model), the *z* attribute can represent any other continuously varying attribute or *regionalized variable*, such as levels of pollutants in soil, atmospheric pressure, annual precipitation, an index of marketing potential, population density, or the costs of access to a given location.

Basic operations for spatial analysis with discretized continuous fields

MAP ALGEBRA AND CARTOGRAPHIC MODELLING

The Kisii land evaluation examples given in the previous chapter demonstrated that when one has to use data from several geographically overlapping entities it is much easier to maintain the database and to compute new attribute values if all the data are referenced to a uniform geometry, namely that of the regular square grid. The loss of information due to rasterizing smooth polygon boundaries is more than offset by the advantage of not having to create new polygons by intersection. Moreover, by choosing a grid size to match that used by remotely sensed imagery, one has the added advantage that the satellite data can also be used as input for data analysis and modelling. As computers increase in power and data sources improve in resolution (cf. Plate 1) the degradation of a spatial pattern by rasterizing becomes more acceptable than the creation of large numbers of topologically linked lines and polygons.

A further, major advantage with raster representation in which each attribute is recorded in a separate overlay, is that any mathematical operation performed on one or more attributes for the same cell can easily be applied to all cells in the overlay. This means that one can use exactly the same algebraic notation to operate on gridded data as on single numbers. The method is called *Map Algebra* (Tomlin 1983, 1990) and the procedure of using algebraic techniques to build models for spatial analysis is called *Cartographic Modeling* (*sic*).

The methods of map algebra mean that the user needs only to specify the spatial operations to be used and the names of the source overlays and the result—the computer program then applies the operation to all the cells in the overlays. This makes it very easy to write computer models as sequences of computations, and makes the extension of formerly point models to two-dimensional space very easy. For example, the command:

$$\text{NEWMAP} = \text{MAP1} + \text{MAP2} + \text{MAP3} \tag{8.1}$$

is all that is necessary to compute the sum of the values of the attributes on the three overlays called MAP1, MAP2, and MAP3. The command:

$$\text{NEWMAP} = (\text{MAP1} + \text{MAP2} + \text{MAP3})/3 \tag{8.2}$$

computes the average value, and the command:

$$\text{NEWMAP} = (\text{SQRT}(\text{MAP1}) + \text{SQRT}(\text{MAP2}) + \text{SQRT}(\text{MAP3}))^{**}2 \tag{8.3}$$

computes the squared sum of square roots.

All these examples compute new values on a cell-by-cell basis: they are known as *point operations* and are formally equivalent to the same mathematical operations that were applied to the attributes of single point, line, or polygon entities in Box 7.1. The extra advantages of the grid-based approach become clear when we see that by using concepts of surface differentiation and smoothing it is possible to compute attributes that are some *spatial function* of the area or neighbourhood surrounding a given cell. As we explain in this chapter, there are many useful neighbourhood functions which can be used for spatial analysis. Finally, because the grid-based approach is an approximation of a continuous surface it is possible to determine new attributes such as views over the surface, or to extract the topology of the steepest downhill paths. These operations provide a rich toolkit for studying phenomena as varied as cross-country visibility and

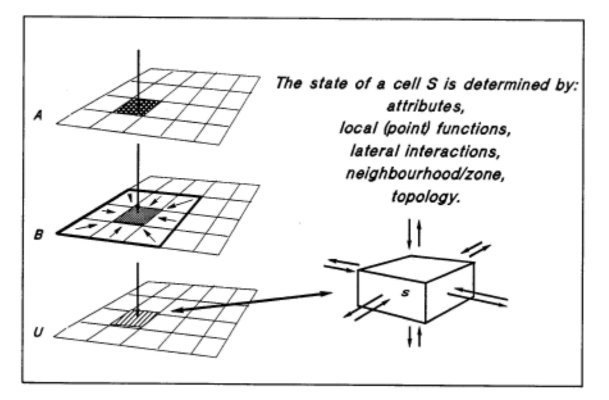


Figure 8.1. The state of a cell is a function of local operations and lateral interactions with its neighbours

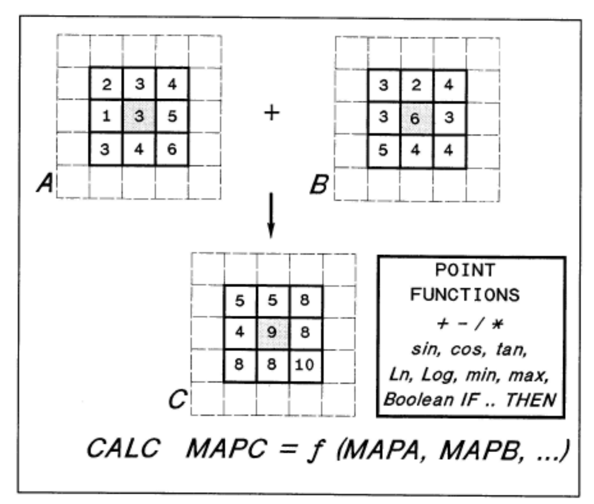


Figure 8.2. Examples of point operations on a cell

hydrological processes or for optimizing access to particular kinds of terrain.

From the foregoing, we see that if the command language interface (CLI) allows the user to express the basic spatial functions in a mathematical language then it will be easy to write mathematical models to operate on the gridded data. Many GIS provide a simple programming language called a *macro language* for this purpose. In this book we assume that all operations can be expressed in simple, general, mathematical terms so that the generic model code used to explain the examples can easily be implemented on various systems. Not all authors of Map Algebra are happy with mathematical formulations, however, and Tomlin (1990) provides an English-language alternative for those who cannot cope with the mathematical conventions. It is our belief, however, that the use of conventional mathematical terminology is much preferable.

In this chapter we review the main kinds of mathematical operation that can be used with gridded data, and illustrate the operations with a range of simple and more complex practical examples.

Point operations

All the logical and numerical operations presented in the previous chapter for simple point, line area entities linked to a RDBMS can be used on the individual grid cells of discretized continuous fields. This means that the values for the same grid cell in different overlays can be logically selected, added, subtracted, or manipulated by any mathematical method that is permitted for the data type in question (Figure 8.1 and 8.2). Therefore one can write commands that can add up or subtract real numbers, but not numbers that are coded as a Boolean or nominal data type. Boolean or nominal data can be operated on by logical commands in the same ways as given in the previous chapter.

Spatial operations

Using gridded data has advantages and disadvantages compared with a topologically linked vector database of defined entities. The disadvantages include the problems that the exact shapes of entities are only approximated by the grid cells and that directed operations over a network cannot be carried out without first deriving the topology from the properties of the surface. The advantages are that the continuous field model provides a much richer suite of truly spatial analysis operations that have many practical uses.

The following operations compute a new attribute for a given cell as some function of the attributes of cells within a certain spatial neighbourhood. The neighbourhood is often, but does not have to be isomorphic (i.e. square or circular). In most cases in GIS the cell size is fixed and uniform over the whole domain of interest. The adoption of grids of variable density would require modifications to the algorithms but would not essentially change the character of the operations being carried out. These spatial operations include:

- Interpolation
- Spatial filtering
- First and higher-order derivatives
- The derivation of surface topology: drainage networks and catchment delineation
- Contiguity assessment (clumping)
- Non linear dilation (spreading with friction)
- Viewsheds, Shaded relief, and irradiance.

Interpolation

Interpolation is the prediction of a value of an attribute $\hat{z}$ at an unsampled site ($\mathbf{x}_0$) from measurements made at other sites $\mathbf{x}_i$ falling within a given neighbourhood. As shown in Chapters 5 and 6, interpolation is used to *create* discretized continuous surfaces from observations at sparsely located points or for resampling a grid to a different density or orientation as in remote sensing images. Interpolation can be seen as a particular class of spatial filtering where the input data are not necessarily already located on a continuous grid. All other methods discussed here assume that the grid has already been created. Remember that the *range* of the variogram can be used to define the radius of the interpolation search radius (Chapter 6).

Interpolation is often a complicated operation and while interpolation operations could in principle be expressed in a mathematical command language most users will encounter specialist packages so that standard terminology cannot be used.

Spatial analysis using square windows

SPATIAL FILTERING

The simplest and perhaps most widely used method of spatial filtering a discretized, continuous surface involves passing a square window (otherwise known as a kernel or filter) over the surface and computing a new value of the central cell of the window $C_{i,j}$ as a function of the cell values covered by the window. This kind of operation is also commonly known as *convolution*. The window is frequently of size 3 × 3 cells, but any other kind of square window (5 × 5, 7 × 7 cells, or distance measurements) is possible. The general equation is:

$$C_{i,j} = f\left(\sum_{i-m}^{i+m}\sum_{j-n}^{j+n} c_{i,j} \cdot \lambda_{i,j}\right) \qquad 8.4$$

where f stands for a given window operator on windows of sides $2m+1, 2n+1$, and λ_{ij} is a weighting factor.

$c_{i-1,j-1}$	$c_{i,j-1}$	${}_{i+1,j-1}c$
$c_{i-1,j}$	$C_{i,j}$	$c_{i+1,j}$
$c_{i-1,j+1}$	$c_{i,j+1}$	$c_{i+1,j+1}$

The most commonly used window operations (f) are low- and high-pass filters.

Smoothing (low-pass) filter The value for the cell at the centre of the window is computed as a simple arithmetic average of the values of the other cells (Figure 8.3). In systems capable of using real numbers the mean can be computed as a real number with a decimal component but in many remote sensing systems both the inputs and the outputs are coded as integer numbers so the decimal part is lost by truncation (see Chapter 8).

In remote sensing systems and image analysis the mean values are computed by multiplying the cell values in the window by the $n \times m$ values in the filter. For example, for a 3 × 3 filter, the mean value for the window centre can be computed by multiplying each cell value by a weight of 1/9 and adding all results. For a 5 × 5 window, the weight of each cell is 1/25. Extra weights can be given to the central cell by introducing weights that are non-linear, i.e. those cells closest to the central cell have larger weights than those further away, much like the idea of distance weighting in ordinary interpolation. For example, for a 3 × 3 window:

1	1/15	2/15	1/15
2	2/15	3/15	2/15
3	1/15	2/15	1/15
	1	2	3

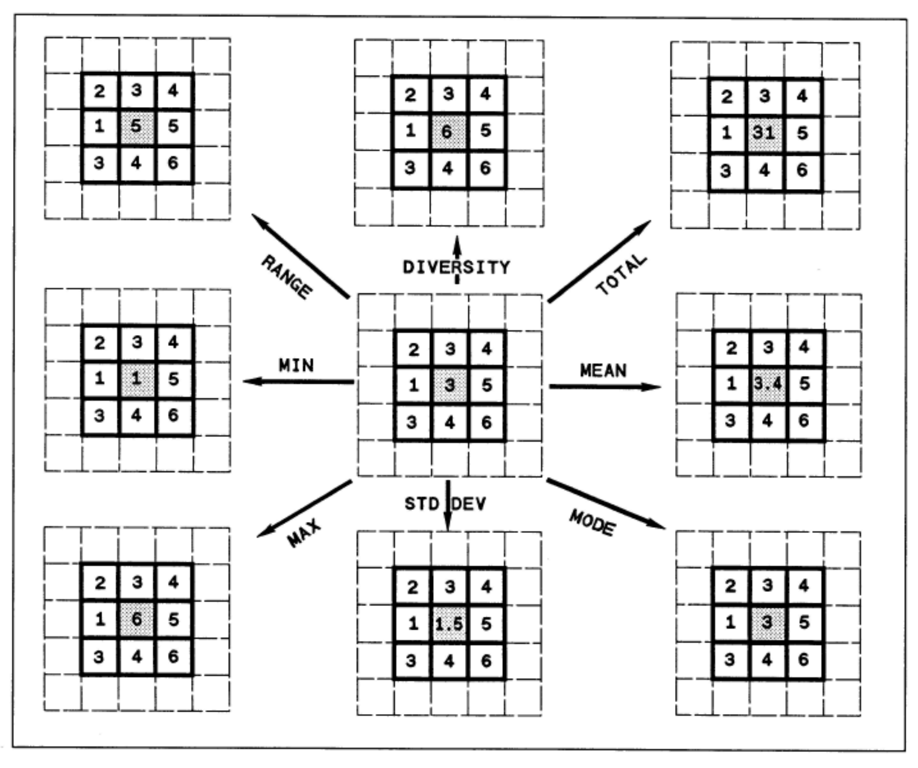


Figure 8.3. Window operations for spatial filtering

or for a 5 × 5 window:

1	1/65	2/65	3/65	2/65	1/65
2	2/65	3/65	4/65	3/65	2/65
3	3/65	4/65	5/65	4/65	3/65
4	2/65	3/65	4/65	3/65	2/65
5	1/65	2/65	3/65	2/65	1/65
	1	2	3	4	5

The low-pass filter has the effect of removing extremes from the data, producing a smoother image (Figures 8.4, 8.5). For nominal and ordinal data (and also to integer and ratio data) the mean can be replaced by the *mode*, which is the most common value (*majority*). Using a modal filter on nominal data (e.g. soil units) can be a useful way of simplifying a complex map (Figure 8.6) but note that smoothing a gridded image with a modal filter is a different kind of operation than the procedure of generalizing a map by reclassifying the attributes and merging the soil polygons given in Chapter 6.

Generic commands for filtering.

To compute low-pass and high-pass filters:
Low-pass = *windowaverage* (continuous_ surface, *n*)
High-pass = continuous_surface — low-pass
where *n* is the side of the square window in cells or distance units
To compute a modal filter:
Modalmap = *windowmajority* (continuous_ surface, *n*)

In a similar way, the local maximum or minimum values, and their difference, the *range* can be easily

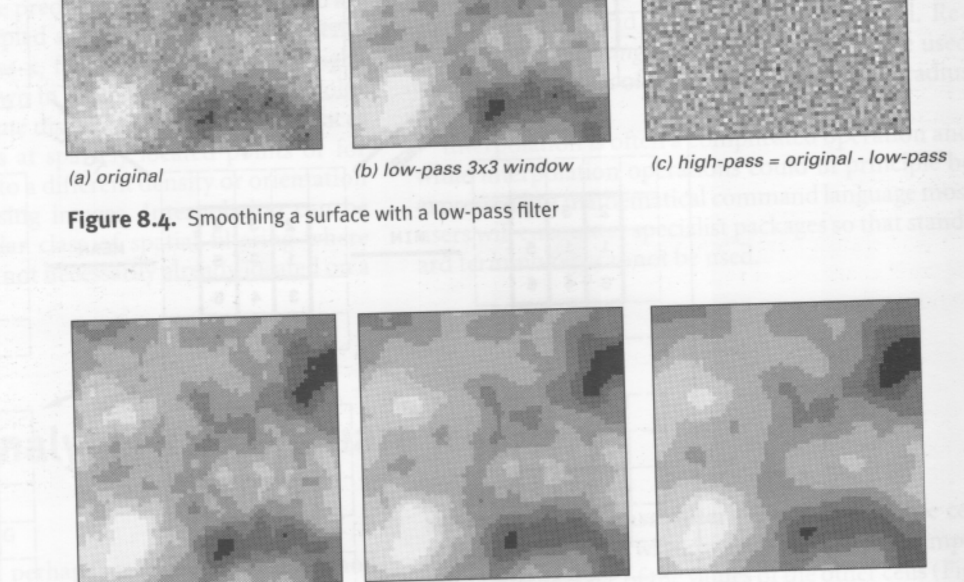


Figure 8.4. Smoothing a surface with a low-pass filter

Figure 8.5. The effect of increasing window size on smoothing

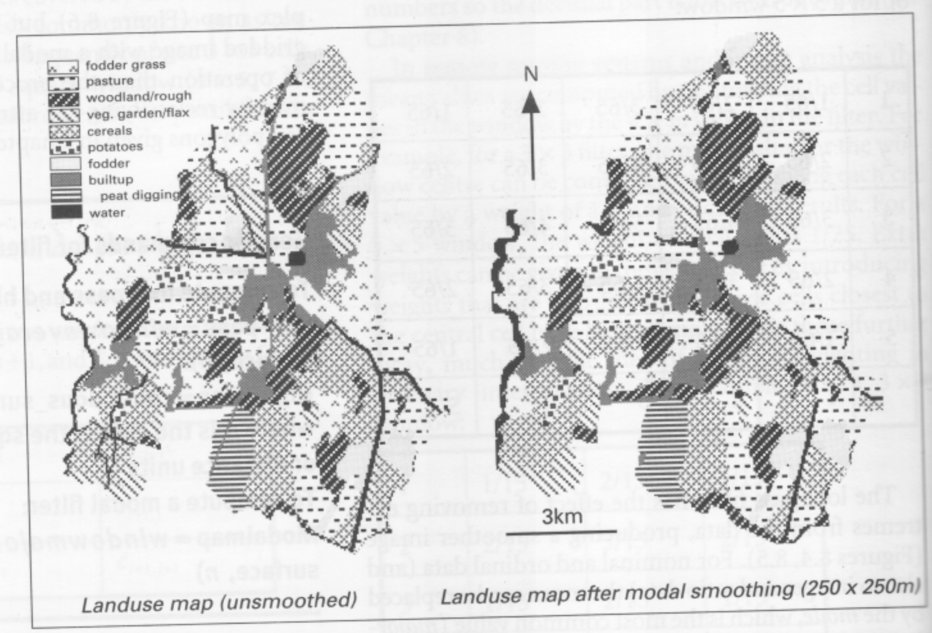


Figure 8.6. Smoothing a complex polygon map with median smoothing aggregates areas but does not reduce the number of classes (cf. Figure 7.5)

computed. Diversity (the number of different values in the window) or the difference between two cells on any one of the four directional axes within the window are alternative options. For nominal and ordinal data the *minority* (the least common) and the *diversity* (the number of different values in the window) are useful operations for indicating the local complexity of the spatial pattern. Each procedure can be applied by using the appropriate operator (see text box).

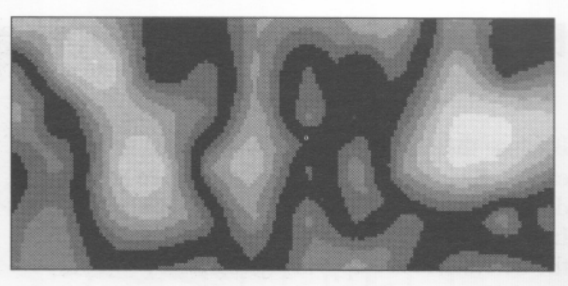
High-pass and edge filters The inverse of the low-pass filter is one that enhances the short range spatial properties of the continuous surface, enhancing areas of rapid change or complexity. The high-pass filter is defined as:

$$\textit{Original surface} - \textit{Low-pass image} = \textit{high-pass image} \quad (8.5)$$

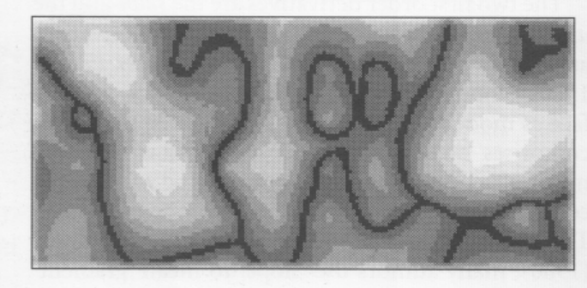
The qualities of the high-pass filter therefore can depend on how the low-pass filter is defined (see above). Alternatively, a set of weights can be defined for the window (Pavlidis 1982). A commonly used set of weights is called the *Laplacian* filter:

0	1	0
1	−4	1
0	1	0

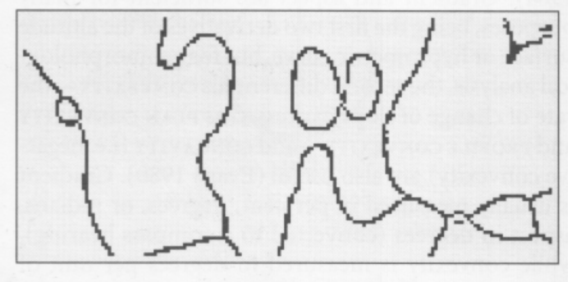
Figure 8.7 shows an example of applying an edge filter to determine the locations of maximum rates of change in a continuous surface. Edge filters are also used to enhance relatively uniform areas in the continuum provided by a remotely sensed image. The derivation of sharp edges and sets of boundary pixels is often used for inferring the presence of discrete spatial entities in the image which ultimately could be extracted and vectorized as required.



(a)



(b)



(c)

Effects of edge filtering to obtain "boundaries"
(a) original surface
(b) high-pass filter
(c) maximum rate of change yields boundaries

Figure 8.7. Using an edge filter to extract boundaries

First and higher order derivatives of a continuous surface

Because the gridded surface is supposed to be mathematically continuous it is in principle possible to derive the mathematical derivatives at any location. In practice, because the surface has been discretized, the derivatives are approximated either by computing differences within a square filter or by fitting a polynomial to the data within the filter.

The two first order derivatives are the *slope* and the *aspect* of the surface; the two second order derivatives are the *profile convexity* and *plan convexity* (Evans 1980). Slope is defined by a plane tangent to the surface as modelled by the DEM at any given point and comprises two components namely, GRADIENT, the maximum rate of change of altitude, and ASPECT, the compass direction of this maximum rate of change. The terms just used follow the terminology of Evans (1980); many authors use 'slope' to mean 'gradient' as just defined (e.g. Peuker *et al.* 1978, Marks *et al.* 1984) and 'exposure' for 'aspect' (e.g. Marks *et al.* 1984). Gradient and aspect are sufficient for many purposes, being the first two derivatives of the altitude surface or hypsometric curve, but for geomorphological analysis, the second differentials CONVEXITY—the rate of change of slope expressed as PLAN CONVEXITY and PROFILE CONVEXITY—and CONCAVITY (i.e. negative convexity) are also useful (Evans 1980). Gradient is usually measured in per cent, degrees, or radians, aspect in degrees (converted to a compass bearing), while convexity is measured in degrees per unit of distance (e.g. degrees per 100 m).

USING DIRECTIONAL FILTERS TO ESTIMATE SLOPE AND ASPECT

The derivatives of the hypsometric curve are usually computed locally for each cell on the altitude matrix from data within a 3 × 3 cell kernel or 'window' that is successively moved over the map (Figure 8.8). The simplest finite difference estimate of gradient in the x direction at point i,j is the Maximum Downward Gradient).

$$[\delta z/\delta x]_{ij} = \max[(z_{i+1j} - z_{i-1,j})/2]/dx \qquad 8.6$$

where δx is the distance between cell centres. (Note that for comparisons along diagonals the $\sqrt{2}$ correction to δx should be applied!). This estimator has the disadvantage that local errors in terrain elevation contribute quite heavily to errors in slope. A better, much-used second-order finite difference method (Fleming and Hoffer 1979, Ritter 1987, Zevenbergen and Thorne 1987) uses a second order finite difference algorithm fitted to the 4 closest neighbours in the window. This gives the slope by

$$\tan S = [(\delta z/\delta x)^2 + (\delta z/\delta y)^2]^{0.5} \qquad 8.7$$

where z is altitude and x and y are the coordinate axes.

The aspect is given by

$$\tan A = -(\delta z/\delta y)/(\delta z/\delta x) \qquad (-p < A < p) \qquad 8.8$$

Zevenbergen and Thorne (1987) show how these attributes and the convexity and concavity are computed from a six-parameter quadratic equation fitted to the data in the kernel—see Box 8.1.

A third-order finite difference estimator using all eight outer points of the window given by Horn (1981) is:

for the east–west gradient,

$$[\delta z/\delta x] = [(z_{i+1,j+1} + 2z_{i+1,j} + z_{i+1,j-1}) - (z_{i-1,j+1} + 2z_{i-1,j} + z_{i-1,j-1})]/8\delta x \qquad 8.9$$

and for the south–north gradient

$$[\delta z/\delta y] = [(z_{i+1,j+1} + 2z_{i,j+1} + z_{i-1,j+1}) - (z_{i+1,j-1} + 2z_{i,j-1} + z_{i-1,j-1})]/8\delta x \qquad 8.10$$

Alternative methods fit a multiple regression to the nine elevation points in the 3 × 3 window and derive the slope and aspect from that.

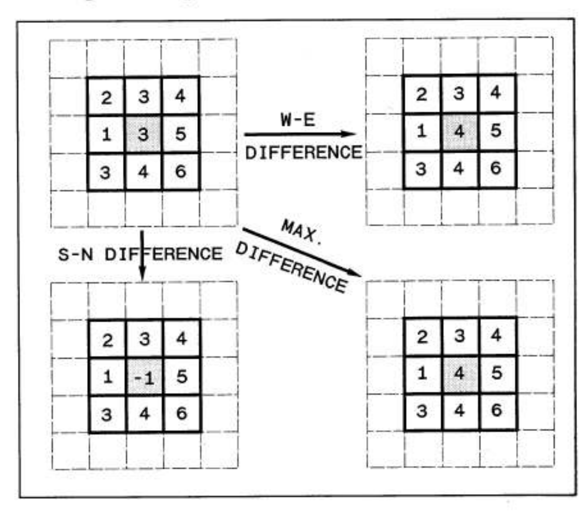
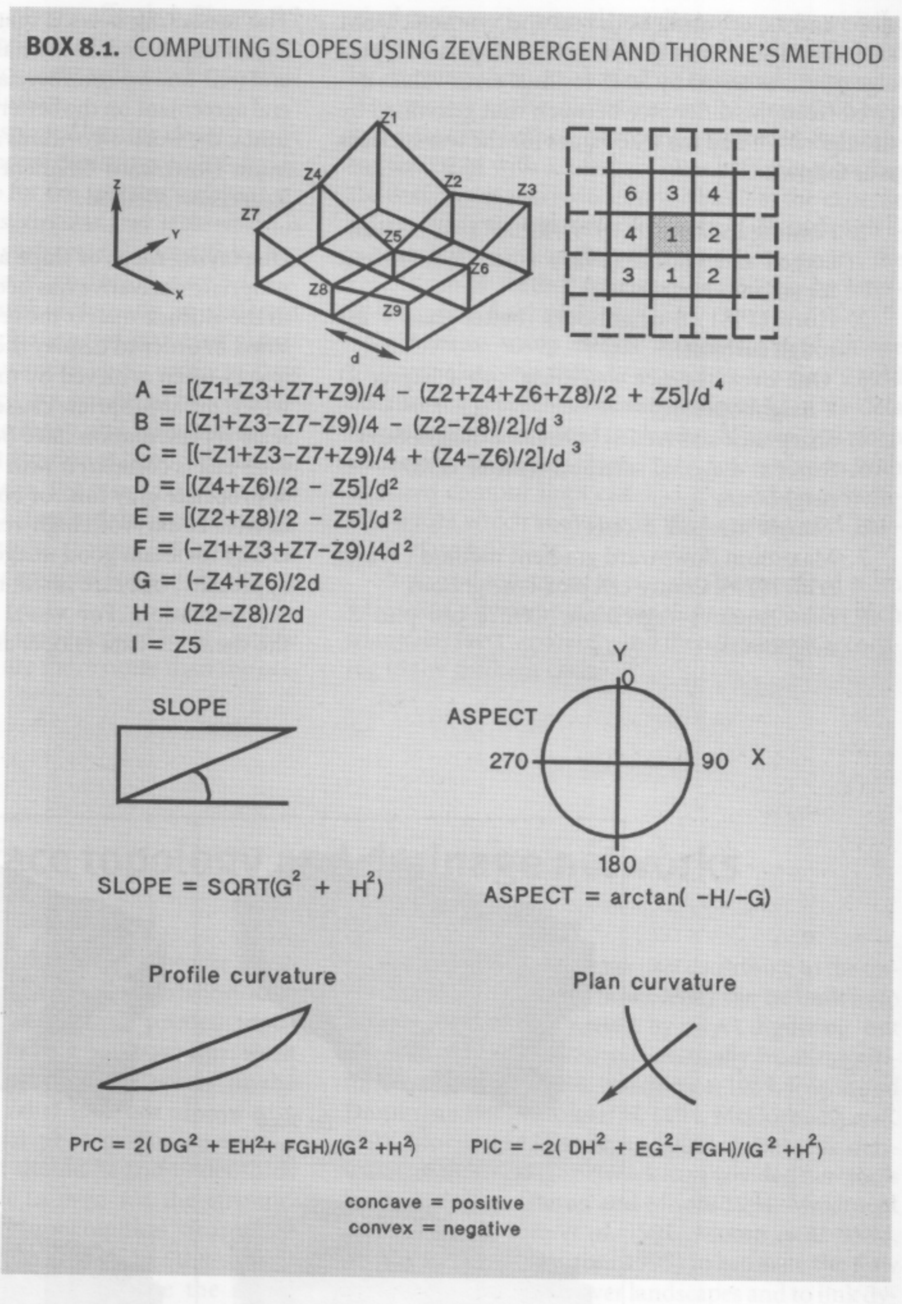


Figure 8.8. Computing derivatives with simple filters

BOX 8.1. COMPUTING SLOPES USING ZEVENBERGEN AND THORNE'S METHOD



Given the variety of methods available for computing slope and aspect it is useful to know which is best. Skidmore (1989) reviewed six methods of estimating slope and aspect, including those given above. He concluded that both the second and third method given above were superior to the simple algorithm given in equation (8.6) but that there was little difference in the results returned by Horn's method and polynomial method. Hodgson (1995) used Morrison's (1971) synthetic test surface to study algorithms using four or eight neighbours and showed that with those data algorithms employing only four neighbouring cells were consistently more accurate at estimating slope and aspect values than eight-neighbour methods such as Horn's. Recently Jones (1997) has carried out another analysis of eight algorithms for computing

slope and aspect using both real and synthetic DEM surfaces. His order of preference for slope and aspect algorithms as tested by RMS residual error values derived from the difference between that generated by the algorithm and the true values for the test surfaces is as follows:

1. Fleming and Hoffer (1979)/Ritter (1987)/Zevenbergen and Thorne (1987) 4-neighbours (best for smooth surfaces)
2. Horn (1981) 8-neighbours (better than 1 for rough surfaces)
3. 'One over distance weighting' (see Chapter 5) 8-neighbours
4. Sharpnack and Atkins (1969)—8-neighbours
5. Ritter's diagonal method (Ritter 1987)—4-neighbours
6. 'Simple method' 3-cells
7. Maximum downward gradient method (Travis *et al.* 1975). Centre cell plus 1-neighbour
8. Least squares regression. Centre cell plus 8-neighbours.

The reassuring news is that Horn's method and the Zevenbergen and Thorne algorithm are used by several well-known commercial GIS, so that there is general agreement on the better algorithms. According to Jones, the worst algorithms are consistently the Maximum Downward Gradient and the Multiple Linear Regression Method.

Displaying maps of slope and aspect After the appropriate derivative has been calculated for each cell in the altitude matrix the results may need to be classified in order to display them clearly on a map. This is very often achieved by means of a lookup table in which the appropriate classes and their colour or grey scale representation have been defined. The value in each cell is compared with the lookup table, and the appropriate grey tone or colour is sent to the output device. Today, with high-resolution display devices it is easy to obtain good images that the human eye can appreciate, but care is still needed to choose the best representation. For visual appreciation, display of the thematic data (slope, aspect, etc.) draped over a

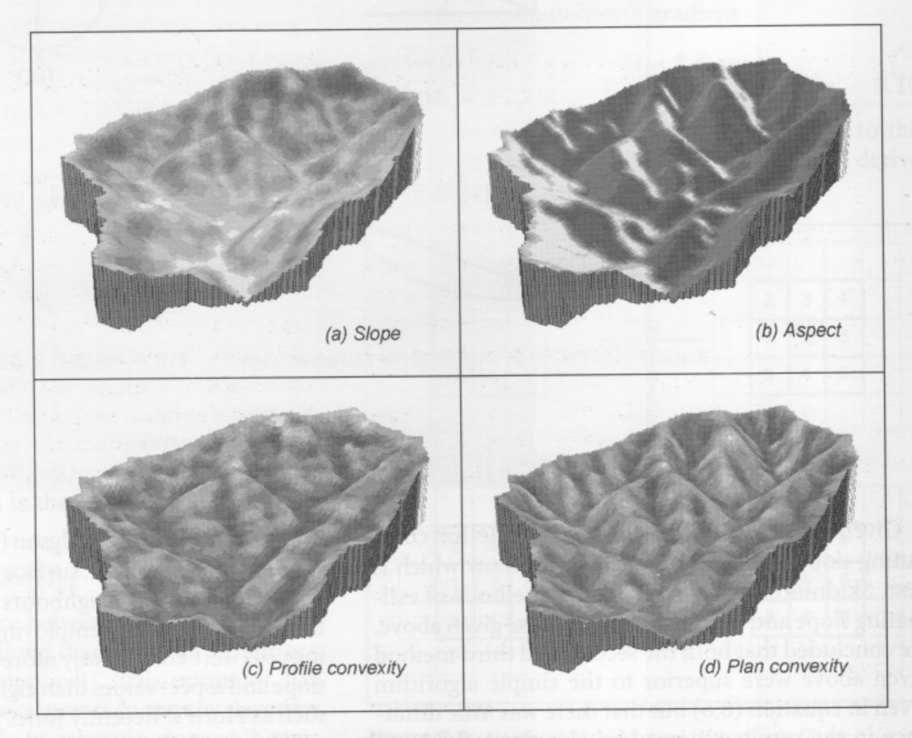


Figure 8.9. First and second order derivatives of a DEM

digital elevation model is very effective. Figure 8.9 gives examples of the slope, profile convexity, and plan convexity for a small catchment with moderate relief having a 30 × 30 m resolution.

Aspect maps can be displayed by nine classes—one for each of the main compass directions N, NE, E, SE, S, SW, W, NW, and one for flat terrain. An alternative is to use a continuous, circular grey scale which is chosen so that NE-facing surfaces are lightest: this gives a realistic impression of a 3D surface (Figure 5.12).

Slope often varies quite differently in different regions and although adherents of standard classification systems usually want to apply uniform class definitions, the best maps are produced by calibrating the class limits to the mean and standard deviation of the frequency distribution at hand. Six classes, with class limits at the mean, the mean ±0.6 standard deviations, the mean ±1.2 standard deviations usually give very satisfactory results (Evans 1980; see also Mitasova *et al.* (1995) for original ways of displaying slope information).

It is a general feature of maps derived from altitude matrices that the images are more noisy than the original surface—in general, roughness increases with the order of the derivative. The images can be improved by drawing them as shaded or coloured maps on a laser plotter or the derivatives can be smoothed by a low-pass filter before the results are plotted. Smoothing the DEM with a low-pass filter before computing the derivatives also reduces noise but is at the expense of removing the extremes from the data and underestimates of slope angles will result. If necessary, the results can be further smoothed for display by interpolating to a finer grid.

In systems using integer arithmetic it is unwise to interpolate the altitude matrix to too fine a grid because the problem of quantization noise in the data (rounding off) may lead to numerical problems when estimating the slopes (Horn 1981). Interpolation from digitized contour lines can also give serious systematic errors which show up in the derivatives (Chapter 5, Plate 4.5).

Slope and aspect maps can also be prepared from TIN DEM's by computing the slope or aspect for each triangular facet separately and then shading it according to the gradient class.

Deriving surface topology and drainage networks

Before drainage basins and drainage networks could be analysed quantitatively, they had to be laboriously copied from aerial photographs or printed topographical maps. Besides being tedious, this work inevitably led to an increase in the errors in the data. In areas of gentle relief it is not always easy to judge by eye on aerial photographs where the boundary of a catchment should be and under thick forest it may be difficult to even see the streams. Even on very detailed topographical maps the drainage network as represented by the drawn blue lines may seriously underestimate the actual pattern of all potential water courses. It could be useful, for example, to be able to separate water-carrying channels from dry channels at different times of the year, with the information about water coming from remote sensing and the channels from a DEM.

Drainage networks and streams, catchments (or watersheds), drainage divides or ridges are important properties of real landscapes that contribute to the understanding of material flows. They can be built in to a TIN or an altitude matrix by direct digitizing, but they can also be derived automatically from the altitude matrix (Band 1986, Hutchinson 1989, Jenson and Domingue 1988, Marks *et al.* 1984, McCormack *et al.* 1993, Morris and Heerdegen 1988). Automatic derivation of the drainage network has provided new tools for hydrologists (Beven and Moore 1994, Maidment 1993, 1995, Moore *et al.* 1991, Moore *et al.* 1993, Moore 1996, van Deursen 1995) to estimate the flow of water and sediment over landscapes and to link dynamic models of hydrological processes to GIS.

The following steps are required when automatically deriving drainage networks from altitude matrices.

1. Determine routing The flow of material over a gridded surface is determined by considering the direction of steepest downhill descent. There are several algorithms for calculating this, called the D8 or 8-

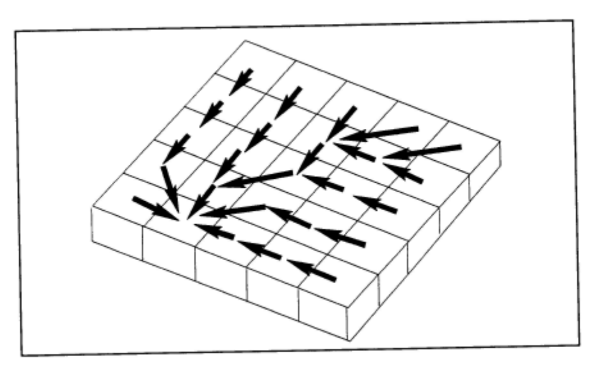


Figure 8.10. Local drain direction vectors to indicate steepest downhill path

point pour algorithm, slope-weighted algorithms, and stream-tube algorithms (Moore 1996).

The D8 (deterministic) algorithm approximates the flow direction by the direction of steepest downhill slope within a 3 × 3 window of cells. This leads automatically to a discretization of flow directions to units of 45°, which is seen by some authors as a serious deficiency. The D8 algorithm computes a new attribute of flow direction which can take eight different directional values which can be expressed as degrees or as numeric codes. A useful implementation is to indicate the directions by the numbers given on the numeric pad of a computer keyboard, so that a *sink* or *pit* is indicated by a '5', but other direction numbering conventions are used (Moore 1996).

7	8	9
4	5	6
1	2	3

The resulting new grid overlay is called the set of *local drain directions* or *ldd*. (Figure 8.10). Each cell contains a *directional* type integer of value *FD* (flow direction) where:

$$FD = d \text{ where } d = f \text{ for } \max_{(f=1,8)} [w_f \,|\, z_{ij} - z_{i,i\pm1,j,j\pm1} \,|\,] \qquad 8.11$$

The distance weight w_f is 1 for NSEW neighbours and $1/\sqrt{2}$ for diagonals.

Figure 8.11*b* is an example of the ldd map displayed over the background of the DEM. Because of its simplicity the D8 algorithm has been incorporated in several commercial GIS. On uniformly sloping surfaces it produces long, linear flow lines, and uniform flow directions, and it is not uncommon to get parallel flow lines that do not converge. It cannot model flow dispersion.

The Rho8 (random) algorithm is a statistical version of the D8 algorithm which was introduced better to represent the stochastic aspects of terrain. It replaces the w_f of $1/\sqrt{2}$ for diagonals by $1/(2-r)$ where r is a uniformly distributed random variable between 0 and 1. Moore (1996) claims that this simulates more realistic flow networks, though like the D8 algorithm it cannot model dispersion (see Desmet and Govers 1966). A D8-based alternative to the Rho8 which might be even more realistic can be obtained by Monte Carlo simulation. An RMS error can be added to the DEM and the D8 algorithm is used to compute a network which is stored. This is repeated 100 times to yield a most probable network (see Chapter 10). The extra advantage of the Monte Carlo simulation is that the error on the DEM can be adjusted to realistic levels and probabilistic flow paths are generated.

FD8 and FRho8 algorithms. These are modifications of the original algorithms allowing flow dispersion or catchment dispersion to be modelled. Flow can be distributed to multiple nearest-neighbour nodes in situations where there is overland flow, rather than concentration of flow in channels, where the D8/Rho8 algorithms are used. The proportion of flow to the multiple downstream nodes is computed on a slope-weighted basis (Freeman 1991, Quinn *et al.* 1991).

Stream tube methods. Costa-Cabral and Burges (1993) determine the amount of flow as a fraction of the area of the source pixel entering each pixel downstream as determined by the intersection of a line indicating the drainage direction (aspect) and the edge of the pixel. See Mitasova and Hofierka 1993 (cited in Moore 1996).

2. Removal of pits When a smooth continuous surface is approximated by a square grid it is inevitable that some cells will be surrounded by neighbours that all have higher elevations. These *pits* could be real closed depressions or merely artefacts of the gridding process. Pits that are artefacts are often generated in narrow valleys where the width of the valley bottom is smaller than the cell size and they can occur at all levels of resolution. They can also occur in areas of gentle relief through errors in interpolation (e.g. Figure 8.11a).

The problem with artefact pits is that they disrupt the drainage topology and need to be removed to obtain a continuous ldd net. They can be removed by one

of two strategies: cutting through or filling up. Cutting through one or more layers of boundary cells to find the next downstream cell requires enlarging the size of the search window to find a cell or series of cells of the same elevation or lower as the core (pit) cell. Once this path has been found the appropriate topological links are written into the cells along the path, irrespective of their true elevation. Filling up involves increasing the elevation of the core cell until it is equal to one or more of its neighbours, and then examining if the neighbour drains downhill to another destination. If this does not happen the elevation is increased again until a linkage is found.

Pit removing is an interactive process regarded by some as a necessary evil. Hutchinson (1989) has developed a spline interpolator for ensuring that pits do not occur, but it is not always sensible to remove all pits automatically because closed and semi-closed depressions may be real features in some landscapes. MacMillan *et al.* (1993) give a detailed description of an application in Alberta, Canada, where the study of surface water storage after the spring snow melt depends on the presence of many partially interconnected depressions. Van Deursen (1995) provides a method for terminating the pit removal process for given values of depression volume, area, or depth and in terms of the received precipitation (i.e. for some levels of run-off the depression behaves as a pit and for others it is full of water and is part of the channel); McCormack *et al.* (1993) provide a feature-based approach for assigning drainage directions on plateaux and depressions.

Once pits and plateaux have been identified then the DEM can be adjusted so that they have been removed. For large and complex data sets with large area pixels a practical alternative is to digitize the river network, convert the river vectors to grid cells, and then 'burn in' the river cells at a lower level in the drainage network.

Example of a generic command for extracting surface topology from a gridded DEM.

lddmap = (dem.map, a, p1, p2, p3, p4, . . .)

where lddmap is the derived topology, a is the algorithm used, and p1, p2, p3, . . . are parameters for removing pits according to their outflow depth, core volume, core area, etc.

Using the ldd network for spatial analysis

Irrespective of the algorithm used to compute the flow directions, the result is to create a gridded overlay in which the surface topology has been made explicit (e.g. Figure 8.11*b*). This ldd network is extremely useful for computing other properties of a digital elevation model because it explicitly contains information about the connectivity of different cells. This knowledge makes it possible to address problems of directed flow and the transfer of fluids or material without recourse to ad hoc search windows.

Accumulating fluxes of material over a net

Because in a topologically correct network, each cell is linked to a downstream neighbour, it is very easy to compute attributes such as the cumulative amount of material that passes through each cell. The *accumulation* operator computes the new state of the cell as the sum of the original cell value plus the sum of the upstream elements draining to the cell

$$S(c_i) = S(c_i) + \sum_{u}^{n}(c_u) \qquad 8.12$$

If the material value for each cell is 1, the result gives the *upstream element map*, or in other words, the cumulative number of cells upstream of the current cell that discharge through that cell. The upstream element map is usually displayed on a logarithmic scale (Figure 8.11*c*—see also Plates 3.7 and 3.8).

If the material value is supplied from another overlay, for example, effective precipitation, then the accumulate operator will compute the cumulative

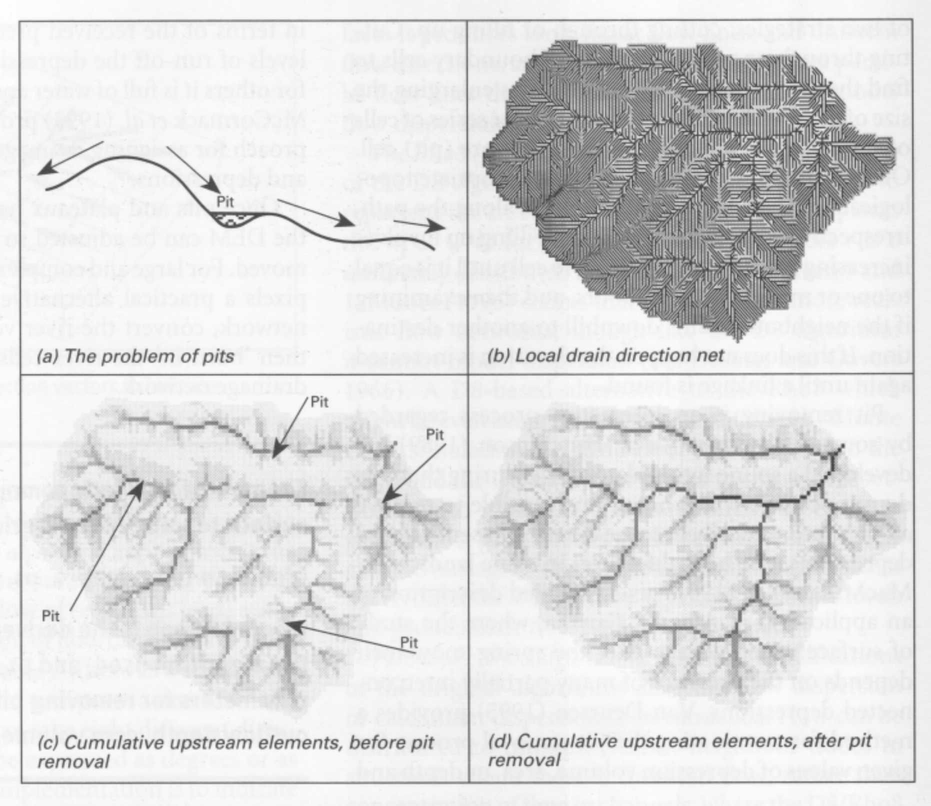


Figure 8.11. Deriving a drainage network from a gridded DEM

flow over an ideal surface. For example, it is easy to compute a mass balance for each cell in terms of

$$S = P - I - F - E \qquad 8.13$$

where S is surplus water per cell, P is input precipitation, I is interception, F is infiltration, and E is evaporation. The cumulative flow over the net is then obtained by accumulating S over the linked cells.

Topological networks are also the basis for a wide range of dynamic modelling tools in GIS, which are explained in more detail in van Deursen and Burrough (1998).

The upstream element map can itself be useful for computing other indices of the terrain. For example a *wetness index map* can be defined as:

$$\text{wetnessindexmap} = \ln(A_s / \tan\beta) \qquad 8.14$$

where A_s is the contributing catchment area in m^2 (number of upstream elements × g the area of each grid cell) and β is the slope measured in degrees (Beven and Kirkby 1979). Figure 8.12*a* shows a wetness map draped over the DEM from which it was derived.

The *Stream Power Index* is defined as

$$\omega = A_s * \tan\beta \qquad 8.15$$

This is directly proportional to the stream power $P = \rho g q \tan\beta$ where ρ is the density of water, g is the acceleration due to gravity, and q is the overland flow discharge per unit width (Moore *et al.* 1993), which is the rate of energy expenditure over time and is a measure of the erosive power of overland flow.

The *Sediment transport Index* is defined as

$$\tau = [A_s/22.13]^{0.6} * [\sin\beta/0.0896]^{1.3} \qquad 8.16$$

This index characterizes the processes of erosion and deposition, in particular the effects of topography on soil loss; it resembles the length-slope factor of the Universal Soil Loss Equation (Wischmeier and Smith 1978) but is applicable to three-dimensional surfaces. Figure 8.12*b* shows that the sediment transport index can vary along the length of a stream.

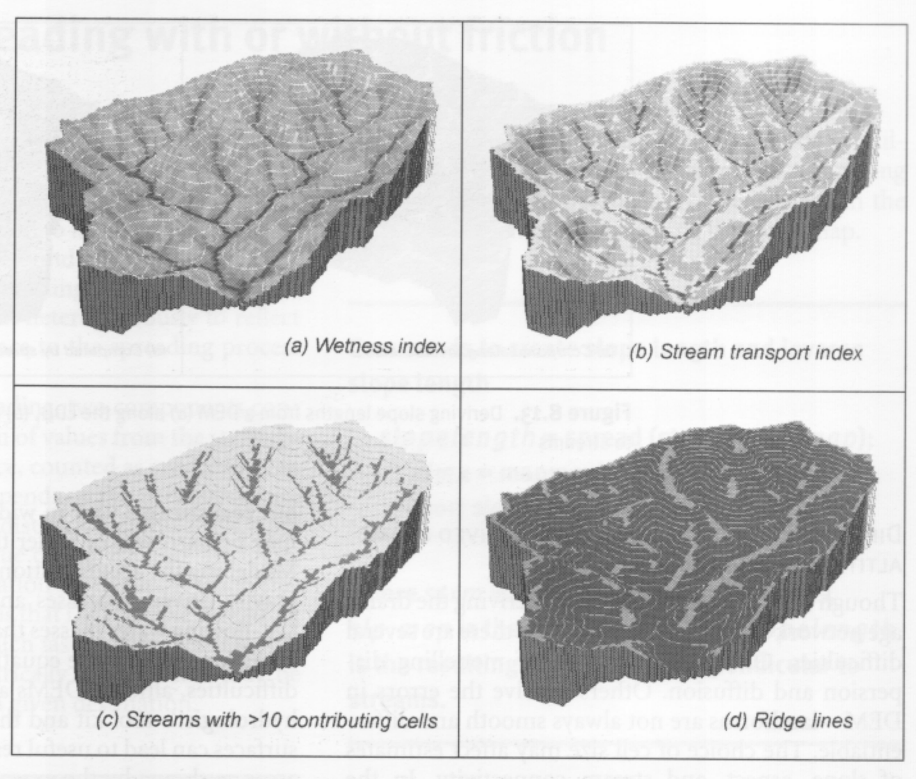


Figure 8.12. Properties derived from the drainage network

OTHER PRODUCTS THAT CAN BE DERIVED FROM THE LOCAL DRAIN DIRECTION MAP

Stream channels can be defined as cells having more than *N* contributing upstream elements. Stream channels can be determined by using a Boolean operator such as:

$$\text{Streams} = \text{if(upstreamelements} \geqslant 50 \text{ then } 1 \text{ else } 0) \qquad 8.18$$

This creates a binary map in which all cells with 50 or more upstream elements are defined as belonging to the set of 'streams' (Figure 8.12*c*).

Ridges. By definition, ridges have no upstream elements, so selecting all cells with an upstream elements value of 1 provides a first estimate of ridges (Figure 8.12*d*).

Catchments. Because all cells that drain through a given cell are part of the catchment of that cell, counting upstream over the ldd automatically computes the area and defines the catchment of the cell. A catchment mask can be computed by assigning a 'one' to all cells in the catchment and a 'zero' to those outside. This can be used as a 'cookie cutter' to identify catchment-specific data from remotely sensed imagery or other sources at the same level of resolution. Using a high-pass or edge filter yields a linear catchment boundary which can be vectorized by converting the cell representation to a chain code (Chapter 3).

The *Slopelength* operator is similar to the *accumulate* operator but it computes a new attribute of a cell as the sum of the original cell value and the upstream cells multiplied by the distance travelled over the network, d_u.

$$S(c_i) = S(c_i) + \sum_u^n (c_u {}^* d_u) \qquad 8.17$$

Figure 8.13*a* shows slope lengths computed in this way. The distance travelled can be a simple Euclidean distance depending on the size of the cells (1 * unit cell size for N–S and E–W; 1.414 for diagonals), or it can include a friction term to deal with resistances within the cells on the network (see Spreading with Friction, below).

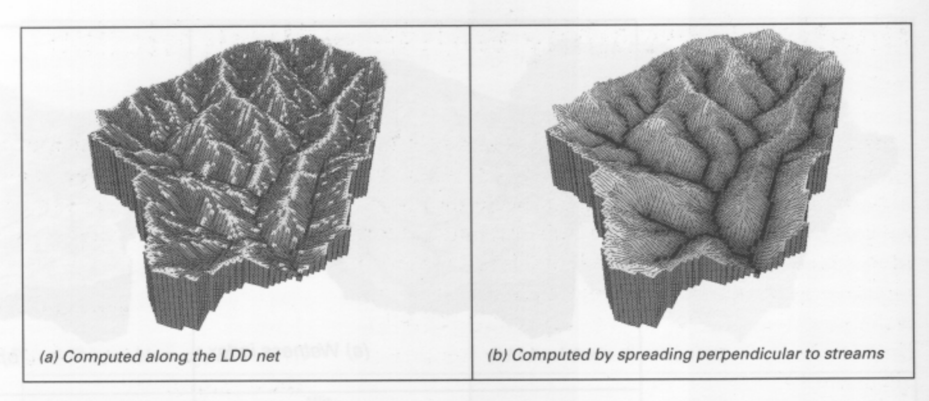


Figure 8.13. Deriving slope lengths from a DEM (*a*) along the LDD, (*b*) by spreading perpendicular to streams

Difficulties with drainage nets derived from altitude matrices

Though there are many benefits of deriving the drainage network from the altitude matrix there are several difficulties. One is the problem of modelling dispersion and diffusion. Others involve the errors in DEM—landforms are not always smooth and differentiable. The choice of cell size may affect estimates of slope, aspect, and stream connectivity. In the altitude matrix the streams are always one cell wide, but real streams vary in width over their length and may be narrower or wider than the cell dimensions. Modelling the accumulation of flows assumes simple gravity driven processes and ignores the inertia of fast-flowing water masses that need to be approached using kinematic wave equations. In spite of all these difficulties, altitude DEMs are finding a place in the hydrologist's tool kit and the analysis of continuous surfaces can lead to useful results in other application areas as shown by the examples given in the second half of this chapter.

Clumping

Very often the result of a Boolean selection or classification on the attributes of cells will result in sets of cells that are spatially contiguous but which cannot be identified as being part of a spatial 'entity'. The *clump* operator examines every cell to see if any of its immediate neighbours in a 3 × 3 window have the same class—if so, then both cells are assigned to the same clump and given a value that identifies that clump as distinct from others. The result is that each contiguous group of cells is aggregated into a larger spatial unit, which could be useful for many purposes. For example, identifying all 'ridges' via the upstream element map may create several loose aggregations of cells that belong to different ridges. Applying the clump operator will identify each cell with a specific clump.

Dilation/spreading with or without friction

This is not a window operation, but a continuous analogue of the dilation or buffering operations on exact entities. Whereas dilation (or buffering) of exact entities is usually limited to isotropic and isomorphic spreading (a buffer around a circle is just a larger circle—Figure 8.14), spreading over a continuous surface can be carried out heterogeneously to reflect the variations in resistance to the spreading process (Figure 8.15).

In non-isotropic spreading, two components contribute to the cumulation of values from the starting-point. The first is distance, counted as cell steps or in real units. The second depends on the attributes of the cells through which the distance accumulation takes place. The larger the value of the 'friction' attribute, the greater the accumulation of 'distance' when traversing a cell. The result is that the effective spreading distance accumulates much faster where resistance is greatest, so that geometrically longer paths may be 'cheaper' ways to reach a given destination.

Operators for spreading with friction.

Spreadmap = spread(startingpoints, v. friction)

where starting-points gives the locations (cells) from which to start the spreading or buffering, v is an initial value, and friction gives the internal resistance on a cell-by-cell basis

---

Both simple and frictional spreading can be used to estimate slope lengths perpendicular to the stream nets derived from the upstream element maps (or from any other linear feature such as roads or railways). Figure 8.13*b* gives an example of computing 'distance from the stream' as a buffer for which the friction has been computed using the slope map.

Commands to create slope length and inverse slope length

***slopelength* = spread (strm, 0, *slp.map*);**
***slmx* = mapmaximum(*slopelength*);**
report slopelength = ((1-(*slopelength*/*slmx*)) * *slmx*);

where *strm* is the map of stream locations, *slp.map* is the map of slopes and *slopelength* is the resulting slope length perpendicular to streams.

Non-isotropic spreading yields a pit-free continuous surface. The 'drainage network' over such a surface defines the set of optimal paths from each cell to the starting-point. Computing the 'catchments' and

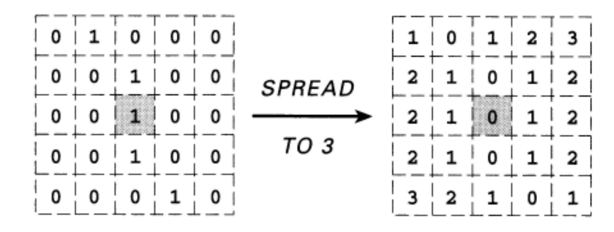


Figure 8.14. Isotropic spreading with grids

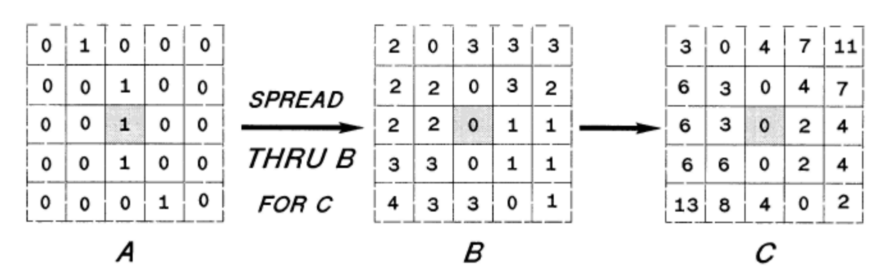


Figure 8.15. Spreading through a resistance function

upstream elements for such a surface can indicate the best routes to follow.

Instead of spreading over all cells, non-isotropic spreading can be confined to follow the routes defined by the ldd map, or by a subset of the ldd. When the resistance varies over the network this can reveal areas where flow problems might accumulate.

Viewsheds, shaded relief, and irradiance

Three, strongly related methods concern the computation of the paths of light between a light source on or above the DEM and its effect at other locations. Whereas all the spatial analysis methods discussed so far concern the attribute value of cells, or the differences in attribute values between cells in the plane of the map overlay, the following operators are concerned with establishing new attributes that refer to the three-dimensional form of the continuous surface. The methods are for the computation of line of sight (determining the viewshed), for computing surfaces with shaded relief for quasi-3D display, and for computing the diurnal or annual inputs of solar energy.

Line of sight maps

This is the simplest operation; the aim is to determine those parts of the landscape that can be seen from a given point. Intervisibility is often coded as a binary variable—0 invisible, 1 visible. The collective distribution of all the 'true' points is called the *viewshed.* Over large distances it is necessary to take the curvature of the earth into account and also the transparency of the atmosphere may be important.

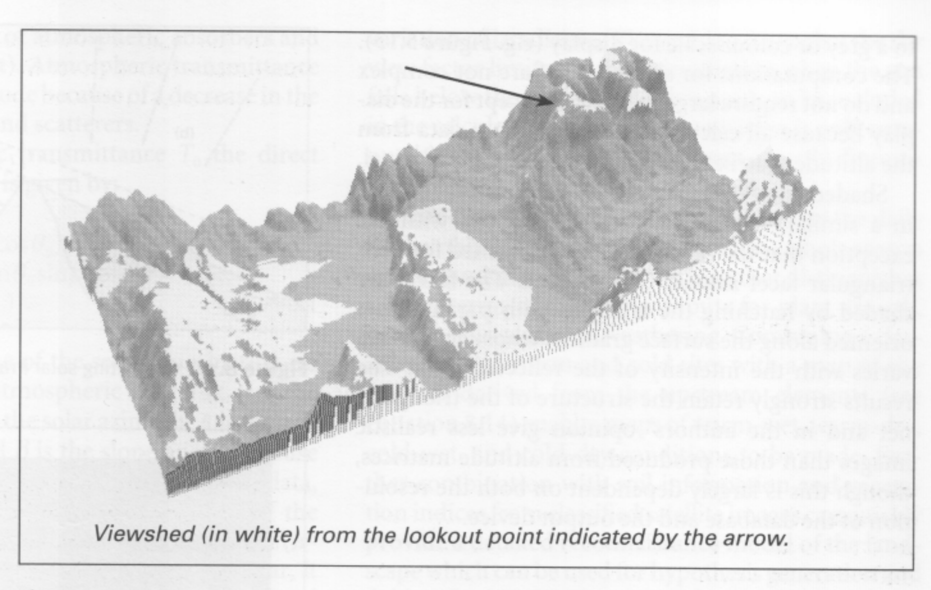
Determining intervisibility from conventional contour maps is not easy because of the large number of profiles that must be extracted and compared. Intervisibility maps can be prepared from altitude matrices and TIN's using tracking procedures that are variants of the hidden line algorithms already mentioned. The site from which the viewshed needs to be calculated is identified on the DEM and rays are sent out from this point to all points in the model. Points (cells) that are found not to be hidden by other cells are coded accordingly to give a simple map (Figure 8.16). Because DEM's are often encoded directly from aerial photographs, the heights recorded may not take into account features such as woods or buildings in the true landform and the results may need to be interpreted with care. In some cases the heights of landscape elements may be built into a DEM in order to model their effect on intervisibilty in the landscape. Viewsheds can also be calculated from TINs (De Floriani and Magillo 1994; Lee 1991*b*).

Estimating the intervisibility of sites is an important GIS application in simulators for pilot training, the location of microwave transmission stations, the appreciation of scenery, or the location of forest fire warning stations. The effects of errors in the DEM on the computation of viewshed have been studied by Fisher (1995).

Shaded relief maps

Cartographers have developed many techniques for improving the visual qualities of maps, particularly with respect to portraying relief differences in hilly and mountainous areas. One of the most successful of these is the method of relief shading that was developed largely by the Austrian and Swiss schools of cartography and which has its roots in chiaroscuro, the technique developed by Renaissance artists for using light and shade to portray three-dimensional objects. These hand methods relied on hand shading and airbrush techniques to produce the effect desired; consequently the end-product, though often visually very striking, was very expensive and was very dependent on the skills of the cartographer who, one suspects, was often also something of a mountaineer.

As digital maps became a possibility, many cartographers realized that it might be possible to produce shaded relief maps automatically, accurately, and reproducibly. Horn (1981) has given an extensive and thorough review of the developments and methods that have been tried. The principle of automated



Viewshed (in white) from the lookout point indicated by the arrow.

Figure 8.16. Viewshed from a look-out point draped over a DEM

shaded relief mapping is based on a model of what the terrain might look like were it to be made of an ideal material, illuminated from a given position. The final results (e.g. Chapter 5, Figure 5.12*b*) resemble an aerial photograph because of the use of grey scales and continuous tone techniques for portrayal, but the shaded relief map computed from an altitude matrix differs from aerial photographs in many ways. First, the shaded relief map does not display terrain cover, only the digitized land surface. Second, the light source is usually chosen as being at an angle of 45° above the horizon in the north–west, a position that has much more to do with human faculties for perception than with astronomical reality. Third, the terrain model is usually smoothed and generalized because of the data-gathering process and will not show the fine details present in the aerial photograph.

The shaded relief map can be produced very simply. All that is required are the estimates of the orientation of a given surface element (i.e. the components of slope) and a model of how the surface element will reflect light when illuminated by a light source placed 45° high to the NE. The apparent brightness of a surface element depends largely on its orientation with respect to the light source and also to the material. Glossy surfaces will reflect more light than porous or matt surfaces. Most discussion in the development of computed shaded relief maps seems to have been generated by the problem of how to estimate reflectance (Horn 1981).

According to Horn (1981), the following method is sufficient to generate shaded relief maps of reasonable quality. The first step is to compute the slopes p,q at each cell in the x (east–west) and y (south–north) directions, as given in equations 8.9 and 8.10 above. These values are then converted to a reflectance value using an appropriate 'reflectance map'. This is a graph relating reflectance to the slopes p,q for the given reflectance model used. Horn suggests that the following formulations for reflectance give good results:

$$\text{(i)}\quad R(p,q) = 1/2 + 1/2(p' + a)/b$$
$$\text{where } p' = (p_o p + q_o q)/\sqrt{(p_o^2 + q_o^2)} \qquad 8.19$$

is the slope in the direction away from the light source. For a light source in the 'standard cartographic position' (45° above the horizon in the NW) $p_o = 1/\sqrt{2}$ and $q_o = -1/\sqrt{2}$.

The parameters a and b allow the choice of grey values for horizontal surfaces and the rate of change of grey with surface inclination. $a = 0$, and $b = 1/\sqrt{2}$ are recommended.

$$\text{(ii)}\quad R(p,q) = 1/2 + 1/2(p' + a)/\sqrt{(b^2 + (p' + a)^2)} \qquad 8.20$$

maps all possible slopes in the range 0–1.

Some formulations for reflectance are computationally complex and it may be more efficient to create a lookup table for converting slopes to reflectance. The reflectance value for each cell is then converted

to a grey or colour scale for display (e.g. Figure 5.13). The computations for shaded relief are not complex and do not require large memories, except for the display because all calculations use no more data from the altitude matrix than is necessary to fill the kernel.

Shaded relief maps are produced from TIN DTM's in a similar way to that described above with the exception that the reflectance is determined for each triangular facet instead of every cell. The facets are shaded by hatching the triangles with parallel lines oriented along the surface gradient whose separation varies with the intensity of the reflected light. The results strongly retain the structure of the triangular net and in the authors' opinion give less realistic images than those produced from altitude matrices, though this is largely dependent on both the resolution of the database and the output device.

Applications of shaded relief maps

Shaded relief maps can be extremely useful by themselves for presenting a single image of terrain in which the three-dimensional aspects are accurately portrayed. Not only have they been extremely useful for giving three-dimensional images of the bodies of the solar system, they are finding a growing application in quantitative landform analysis. When used in combination with thematic information they can greatly enhance the realism of the final map in a way that was impossible before computers were available.

Irradiance mapping

This is the extension of the shaded relief principle to compute the amount of solar energy falling directly on a surface. The sun is now not fixed in any one position in sky, but is allowed to take a position according to the latitude, the time of day, and the day of the year. There is a need to incorporate the effect of atmospheric absorption on the amount of energy actually received, and also to model the shadowing effect of terrain (the *sky view factor*), which is of considerable importance in hilly landscapes in winter or at the beginning or end of the day (Dubayah and Rich 1995). Diffuse irradiance is more difficult to calculate as are the exact effects of local reflection.

Detailed information on topographic solar radiation models can be found in Dozier (1980), Dozier and Frew (1990), Dubayah (1992), Kumur *et al.* (1997) and the references cited therein. Direct radiance received at any point is a function of solar zenith angle, solar flux at the top of the atmosphere (exoatmospheric flux), atmospheric transmittance, solar illumination angle, and sky obstruction (Figure 8.17). Zenith angle and exoatmospheric flux vary with time of day and year, and atmospheric transmittance varies

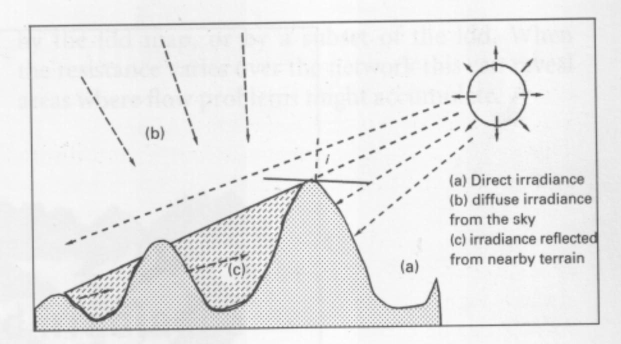


Figure 8.17. Computing solar irradiance for a slope

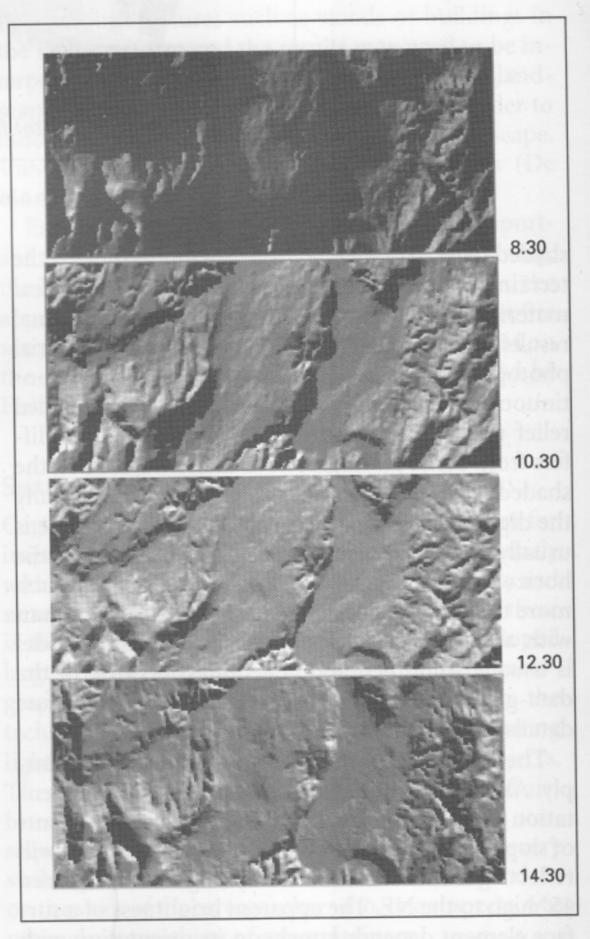


Figure 8.18. Variation in direct irradiance as a function of time of day (view from north)

as a complex function of atmospheric absorbers and scatterers (clouds, dust). Atmospheric transmittance also increases with altitude because of a decrease in the number of absorbers and scatterers.

For an atmospheric transmittance T_0 the direct irradiance I on a slope is given by:

$$\begin{aligned} I &= \cos i S_0 \exp(-T_0/\cos\theta_0) \\ &= [\cos\theta_0 \cos\beta + \sin\theta_0 \sin\beta \cos(\phi_0 - A]S_0 \\ &\quad \times \exp(-T_0/\cos\theta_0) \end{aligned} \qquad 8.21$$

where $\cos i$ is the cosine of the solar illumination on the slope, S_0 is the exoatmospheric solar flux, θ_0 is the solar zenith angle, ϕ_0 is the solar azimuth, A is the azimuth of the slope, and β is the slope angle. Because both β and A are derived from digital elevation data, equation 8.19 describes the spatial variation of the dominant component of incident solar radiation (*irradiance*). As $\cos i$ varies with time of day and year, it is possible to compute both the spatial and temporal variation of irradiance (Figure 8.18). Note that the sky view factor limits the direct irradiation when the sun falls below the horizon and this needs to be included in the calculations (i.e. direct irradiation is only computable for those parts of the terrain directly illuminated or in the 'viewshed' of, the sun).

Note that it is a simple matter to integrate the daily or monthly estimates of irradiance for a whole season or year, and thereby to create a map that distinguishes sites in terms of the energy inputs for plant growth, home heating, or rock weathering. Combining a classified map of warm and cold sites with a map of site wetness derived from the upstream elements (see equation 8.14) enables maps of warm-wet, warm-dry, cold-wet, and cold-dry conditions to be made. Further combination with soil information and vegetation indices from classified satellite images can quickly provide a detailed reconnaissance model of the landscape which can be used for hypothesis generation and fieldwork planning.

Other cell-based analysis operations

The cell-based operations for continuous fields given here are merely the most common subset found in GIS and image analysis. Other operations of similar type have been developed in the areas of *Mathematical Morphology* (see Serra 1968) and *Cellular Automata* (see Gutowitz 1991 and Takeyama and Couclelis 1997).

Temporal change. With gridded data it is very easy to link the operations given in this chapter and also the attribute calculations of Chapter 6 in such a way that they can be carried out many times, thereby providing a means to model dynamic processes (see Van Deursen and Burrough 1998). Most operations with Cellular Automata involve temporal change.

Table 8.1. Summary of attributes that can be computed from DEMs and their application

Attribute	*Definition*	*Applications*
Elevation	Height above mean sea level or local reference	Potential energy determination; climatic variables—pressure, temperature, vegetation and soil trends, material volumes, cut and fill calculations.
Slope	Rate of change of elevation	Steepness of terrain, overland and sub-surface flow, land capability classification, vegetation types, resistance to uphill transport, correction of remotely sensed images
Aspect	Compass direction of steepest downhill slope	Solar irradiance, evapotranspiration, vegetation attributes, correction of remotely sensed images
Profile curvature	Rate of change of slope	Flow acceleration, zones of enhanced erosion/deposition, vegetation, soil and land evaluation indices
Plan curvature	Rate of change of aspect	Converging/diverging flow, soil water properties
Local drain direction (ldd)	Direction of steepest downhill flow	Computing attributes of a catchment as a function of stream topology. Assessing lateral transport of materials over locally defined network.
Upstream elements/area/ Specific catchment area	Number of cells/area upstream of a given cell/upslope area per unit width of contour	Catchment areas upstream of a given location (if the outlet, area of whole catchment), volume of material draining out of catchment
Stream length	Length of longest path along ldd upstream of a given cell	Flow acceleration, erosion rates, sediment yield
Stream channel	Cells with flowing water/cells with more than a given number of upstream elements	Flow intensity, location of flow, erosion/sedimentation
Ridge	Cells with no upstream contributing area	Drainage divides, vegetation studies, soil, erosion, geological analysis, connectivity.
Wetness index	ln(specific catchment area/tan(slope))	Index of moisture retention
Stream power index	specific catchment area * tan(slope)	Measure of the erosive power of overland flow
Sediment transport index (LS Factor)	$(n+1)\left(\frac{A_S}{22.13}\right)^n\left(\frac{\sin\beta}{0.0896}\right)^m$	Characterizes erosion and deposition processes (cf. USLE)
Catchment length	Distance from highest point to outlet	Overland flow attenuation
Viewshed	Zones of intervisibility	Stationing of microwave transmission towers, fire watch towers, hotels, military applications
Irradiance	Amount of solar energy received per unit area	Vegetation and soil studies, evapotranspiration, location of energy-saving buildings, shaded relief

Source: adapted from Moore *et al.* 1993.

Summary of operators that can be used on continuous fields and their products

This chapter has demonstrated that there is a large range of products that can be derived from continuous surfaces that have been discretized as regular grids (Table 8.1). Some derived data, such as slope, aspect, and line of sight, hill shading and irradiance, can also be easily obtained from triangular irregular networks but it is not usual to find systems using TINs that provide facilities for deriving the topology of the surface and for computing material flows over these derived networks. Frequently, the users of TINs will have input their networks explicitly as topologically connected lines or objects.

The most commonly encountered continuous field is the digital elevation model, and most of the derivatives mentioned above have a direct bearing on the use and interpretation of terrain elevation. The operators presented can be used on any continuous field, however, such as remotely sensed images or the results of interpolation or spatial modelling, as will be described in the following examples.

Practical applications of the spatial analysis of continuous surfaces

The rest of this chapter presents several examples that demonstrate the extra value of including spatial operations on continuous fields in GIS analysis. The provision of operators that can be used to analyse spatial interactions and the creation of surface topology opens up many new possibilities, particularly those that involve the transport of materials or fluids over space. One first thinks of hydrology where the movement of water over drainage basins and catchments is an essential part of the process, but there are many other areas where adding spatial interaction to models that previously operated only at points or on an entity-by-entity basis (*lumped models*) provides new predictive power and insights. Note that in this chapter, with the exception of the computation of irradiance, we limit the discussion to static modelling of spatial interactions. Linking time series data with spatial models opens up many more exciting possibilities, but goes beyond the scope of this book. Dynamic modelling in a GIS environment is explained in the companion volume (Burrough and van Deursen 1998).

The areas examined include (*a*) simple hydrological modelling of catchments, (*b*) the modelling of erosion hazards with additional help from remote sensing, (*c*) the optimization of timber extraction from forests, and (*d*) the post-Chernobyl investigation of links between summer flooding and increased ^{137}Cs levels in soils and vegetation in riparian areas of the northern Ukraine. The aim of the examples is to demonstrate how the spatial operators can be used rather than to present fully developed spatial models.

Spatial analysis in surface water hydrology

From the material presented in this chapter it is clear that the analysis of continuous fields provides much information for hydrological modelling, in particular the provision of slope and aspect information, the derivation of drainage nets, the computation of hydrological indices, and the delineation of catchments. Together with other data on potential and actual evaporation, land cover, and soil types, a GIS can provide much input data for running existing surface water hydrological models such as TOPMODEL (Beven *et al.* 1984) and ANSWERS (De Roo *et al.* 1992). Beven and Moore (1994) and Maidment (1993, 1996) provide more details.

Even with a simple raster GIS and a limited command language structure it is possible to carry out

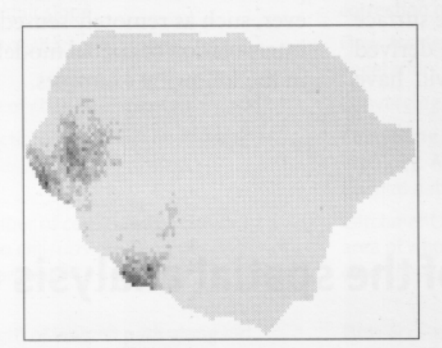
BOX 8.2. COMPUTING MASS BALANCES

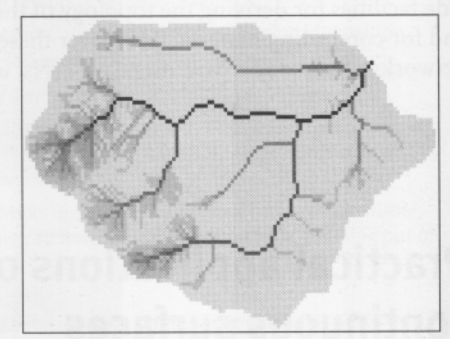
Pseudo code for computing mass balance of moisture over a net

For each cell:

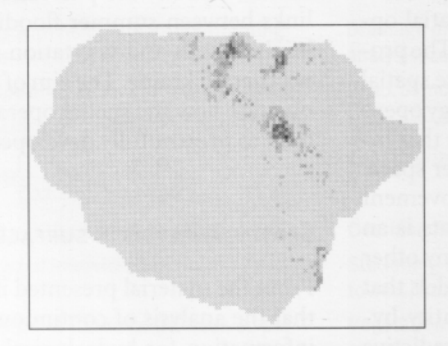
Run-off = (Precipitation – interception – evaporation – storage) + run-on

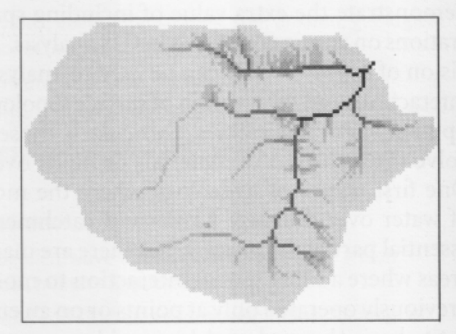
Run-off from each cell occurs when the sum of the water received from upstream cells plus the point water balance exceeds zero; the run-off is added to the next downstream cell.





(a) Effect of precipitation from a simulated storm high in the catchment





(b) Effect of a storm low down in the catchment

Figure 8.19. The effect of a storm on catchment discharge depends on its location within the catchment

simple analysis of hydrological surface processes. Knowing the topological linking between grid cells not only allows material transport from cell to linked cell to be calculated, but also for that transfer to be modified by some other property of the cell. For example, accumulating the water surplus from a mass balance for each cell over the net allows one to model the movement of water surpluses from cell to cell as a flow process (Box 8.2).

Figures 8.19*a*,*b* show how the surplus rainfall from simulated rainfall events high up or low down in the catchment of Figure 8.11 will activate the drainage system down to the catchment outlet. When combined with data on land cover, infiltration, and evaporation,

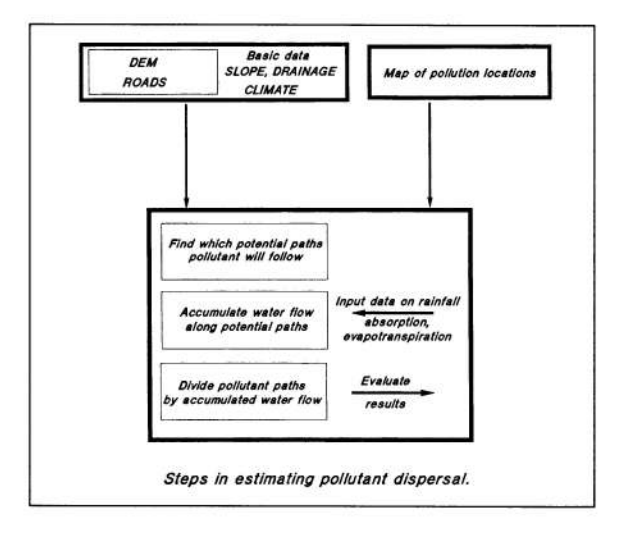
such an analysis can be useful for examining scenarios of rainfall run-off under different kinds of land use.

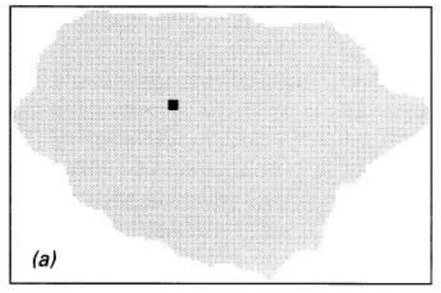
A similar approach allows the easy estimation of pollutant dispersal from point sources if the pollutant is allowed to leak into streams. The procedure is shown in Figure 8.20 as flowchart and example.

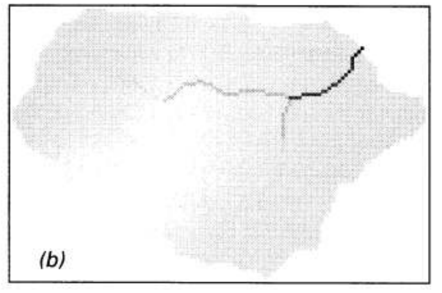
MODELLING SOIL EROSION HAZARDS

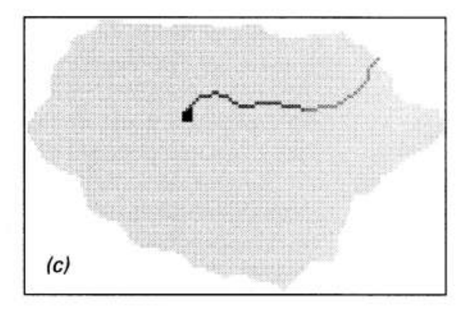
Until relatively recently, soil erosion has been treated by many site managers as a static process, in the sense that each site has been evaluated separately and little attention has been paid to the overland transport of sediment. As shown in the previous chapter, soil erosion models like the USLE (Wischmeier and Smith 1978), SLEMSA (Stocking 1981), or the Morgan, Morgan, and Finney Method (Morgan *et al.* 1984) are site-specific point models. Their calculations ignore both the transport and deposition of eroded soil. Models incorporating overland transport of sediment have been developed but specialists may be needed to run them, particularly if they are written as self-contained modules in FORTRAN or similar programming languages. Also, using site-specific models like the USLE (see Chapter 7) to compute soil erosion over landscapes modelled as sets of independent entities (polygons or pixels) ignores all spatial interactions between the polygons or cells and clearly this approach is physically unrealistic. This aspatial approach to soil erosion has been reinforced by the many field experiments using Wischmeier plots in which the plot of land being studied for its erosion susceptibility has been artificially cut off from the surrounding landscape by a wall of bricks or tiles.

Erosion as a process, rather than as a descriptor, is more properly treated using a data model of continuous variation than as an attribute of crisp entities. The simplest way to introduce sediment transport into









Discharge of pollutants from (a) a point source over (b) the streams leads to (c) dilution

Figure 8.20. Estimating pollutant dispersal with drainage networks

erosion modelling is to treat eroded soil as a material that can flow over the ldd network in a manner similar to water. The *potential erosion* for each cell may be computed using the point models. Then the *transport capacity* of each cell determines how much of the potentially erodible soil can be moved to the neighbouring cell. If the cell has attributes that retard overland flow then little material will be transferred, and if the cell has bare land that cannot offer much resistance to lateral flow, then much or all of the potentially erodible soil will be transported. As each cell is topologically connected to upstream neighbours it will also receive sediment. If the amount of sediment received is greater than that discharged, then deposition occurs; if not, then there is net erosion. The transport capacity of the network depends on geometrical aspects of the landform, such as slope, length of slope, and roughness, and also on the potential and kinetic energy of the water falling on it.

It is fairly easy to add a transport component to the USLE once an ldd network has been established, and the model immediately demonstrates the phenomenon of deposition in valley bottoms, which the point model could never manage. By highlighting combinations of steep slopes, shallow, unprotected soils and aggressive surface flow, the modified models are much better at indicating the location of *erosion potential* over large areas than their point equivalents (cf. Desmet and Govers 1996; Mitasova *et al.* 1996). De Jong (1994) also demonstrated the benefit of using surface topology to improve the prediction of soil erosion hazards from DEMs, remotely sensed land cover data and numerical modelling with his model called *SEMMED*.

SEMMED When assessing erosion hazard for large areas it is useful if data from remote sensing can be used as at least one main input to the model because its continuous variation and resolution in both space and time may more appropriately reflect the nature of spatial and temporal variation than measurements made only at a few point locations that are linked to a priori soil units (polygons). SEMMED uses continuous function analysis to improve the prediction power of erosion models by using topological linkages and data from remote sensing.

SEMMED is based on the Morgan, Morgan, and Finney Method to predict annual soil losses (Morgan *et al.* 1984), modified to take account of overland flow. SEMMED separates the point process of soil erosion into a water phase and a sediment phase. The model considers soil erosion to result from the detachment of soil particles by raindrop impact and the transport of these particles by overland flow, but does not consider splash transport or detachment by run-off.

The water phase uses mean annual rainfall to determine the energy of the rainfall for splash detachment and the volume of run-off. The volume of overland flow is computed using the 'mean rain per rain day' and the 'soil moisture storage capacity'. The water phase does not include exchange of water between the grid cells.

The sediment phase of the model considers soil particle detachment by rainsplash and transport of soil particles by run-off. Detachment is modelled as a function of rainfall energy using the results of an empirical study of Meyer (1981), modified account for rainfall interception by a vegetative cover (Laflen and Colvin 1981). In the sediment phase the transport capacity is calculated as a function of the volume of overland flow, the slope steepness, and the effect of crop cover. USLE C-factor values (Wischmeier and Smith 1978) are used to account for the effects of crop cover.

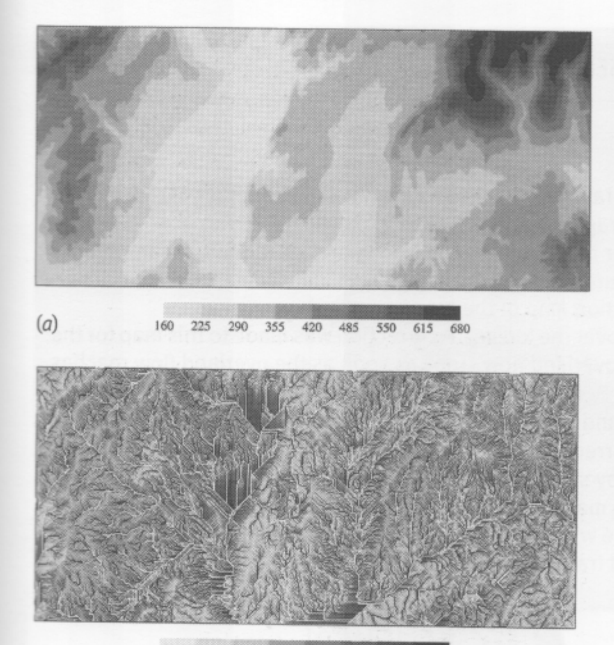
Data sources

Climatological data. A distributed approach to model erosion requires distributed rainfall data, but these are not always available, though radar mapping of rainfall is improving greatly. Lumped climatological data for the Ardèche province were gathered from published sources (CMD 1984). As the multi-temporal satellite images used in this study were acquired in 1984, the annual rainfall figures of that year were used.

Soil physical properties. The French regional soil map (Bornand *et al.* 1977) at a scale 1 : 100 000 was digitized and rasterized into a spatial continuous raster map with a cell size of 30 × 30 m, matching the pixel size of the TM-images. Soil properties required to compute the soil moisture storage capacity were gathered from field data supplemented by literature data and were assigned as attributes to the soil polygon entities of the digital soil map.

Landsat TM. Vegetation data derived from Landsat TM using the NDVI relation (Tucker 1979) were used to compute cell-specific crop cover, evapotranspiration, and interception.

Topographical factors. Contours of the 1 : 25 000 topographical map (IGN 1985) were digitized and interpolated to yield a gridded model DEM (Figure 8.21*a*). Slope steepness was calculated from this DEM using the window procedure given in equation (8.6). The topological network of potential



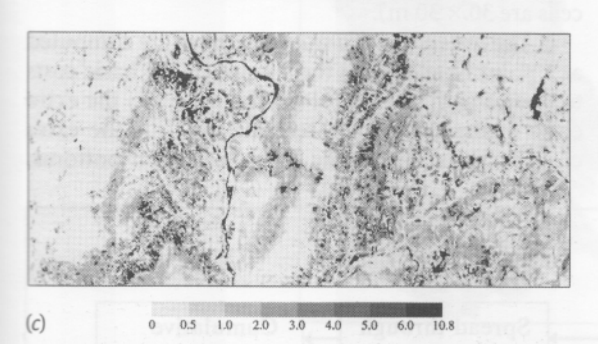


Figure 8.21. (*a*) Digital elevation model of study area for SEMMED (resolution 30 m); (*b*) Transport capacity as computed using the drainage network; (*c*) Erosion hazard (predicted soil loss kg m^{-2} y^{-1})

drainage lines was computed using the 8-point pour algorithm.

The DEM was also used to calculate a distributed map of the transport capacity for SEMMED (Figure 8.21*b*). This map controls the calculations of overland flow once the topsoil moisture capacity per cell is exceeded.

SEMMED results. The predicted rate of soil loss ranged from 0 to 10.16 kg/m^2 (Figure 8.21*c*). Areas with low interception values showed the largest values e.g. the marl areas surrounding the village of Lussas and at the edges of the limestone plateaux. Low values were found in the region of the Massif Central, the higher parts of the Coiron, and the limestone plateaux due to high rainfall interception values by vegetation. SEMMED predicts erosion to occur mainly at the edges of the plateaux (Coiron and the limestone plateaux), which corresponds with field observations and with the presence of gullies and rills as shown on the geomorphological map of the study area. More than 90 per cent of the predicted values of annual soil loss are below 1.5 kg/m^2.

SEMMED demonstrates that it is possible to produce regional maps of erosion hazard, which are much better than simple extrapolations of plot experiments. Once the model has been calibrated it can be used for exploring the effectiveness of measures designed to reduce erosion hazards in the most critical places in the landscape.

Optimal extraction of timber from a natural forest

This example demonstrates how the extraction of timber from a natural forest can be optimized as a function of the costs of access to the forest, the locations of valuable trees, and their market price. The basic data needed include the information necessary to determine the costs of site access, the location of valuable trees, and their current market price.

In this example we assume that the costs of access are determined by the costs of travelling along a single access road and by the costs of traversing land that has no roads. In the latter case we assume that costs of access are determined by the slope of the land and by the costs of crossing stream beds. If wished, we could also include costs incurred by limitations in soil trafficability but the principle is the same.

There are two parts to the analysis, first the determination of the accessibility of each site and thereafter, the determination of which trees will be profitable to extract given certain market conditions.

The idea behind the analysis is that all points of the forest can be reached, but at a price. This price is not the same for every location, but can be modelled as a continuous, cumulative function from the entrance to the forest. This continuous, cumulative function behaves like a DEM, but its Z values are cumulative costs. The shortest paths accross this surface are the optimal access routes from any given cell to the forest entrance.

Figure 8.22 is a flow chart of the operations used. Figure 8.23 shows the basic map data, the digital elevation model, the location of the access road and

BOX 8.3. PROCEDURES USED IN SEMMED

The steps in SEMMED are:

1. Compute mean rainfall per rain day = annual rainfall volume/number of rain days.
2. Compute soil moisture storage capacity for each grid cell.
3. Compute overland flow per raster cell by combining annual rainfall, mean rain per rain day and the soil moisture storage capacity.
4. Derive the local drain direction map of the area.
5. Accumulate overland flow over the ldd net. A correction was made to this map for the river channels: erosion by overland flow stops as soon as the overland flow reaches the river channel.
6. Combine the map of overland flow per rastercell (step 3) with the map of potential overland flow (step 5); a correction was made for the amount of water that infiltrates in the (saturated) top soil by subtracting a map with the saturated infiltration capacity per cell (the infiltration map was produced by reclassifying the digital soil map).
7. Combine the result of step 6 with the slope steepness and the map of the vegetation factor to yield a distributed transport capacity map.

entry point in the south east, and the location of valuable trees. The distribution of trees could have been determined by field survey or by aerial photo interpretation or classification of remotely sensed images.

The steps in the analysis 1. The usual slope and drainage analyses yield maps of slopes and the major streams (in this case more than 100 upstream cells—cells are 30 × 30 m).

2. The *costs of traversing any given cell* are computed as a point operation as the sum of (i) the basic costs to clear a path or drive along the road, (ii) the extra costs incurred by steep terrain, and (iii) the extra costs incurred by having to cross stream bottoms.

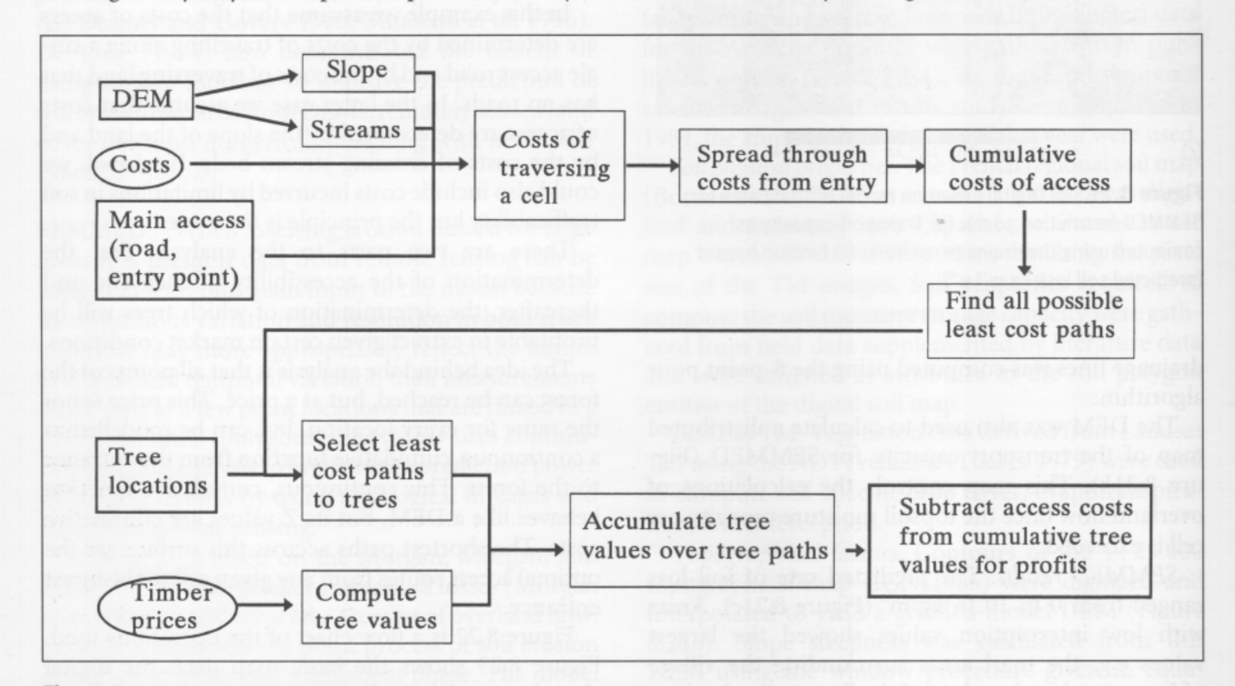


Figure 8.22. Flow chart for optimizing timber extraction from a forest

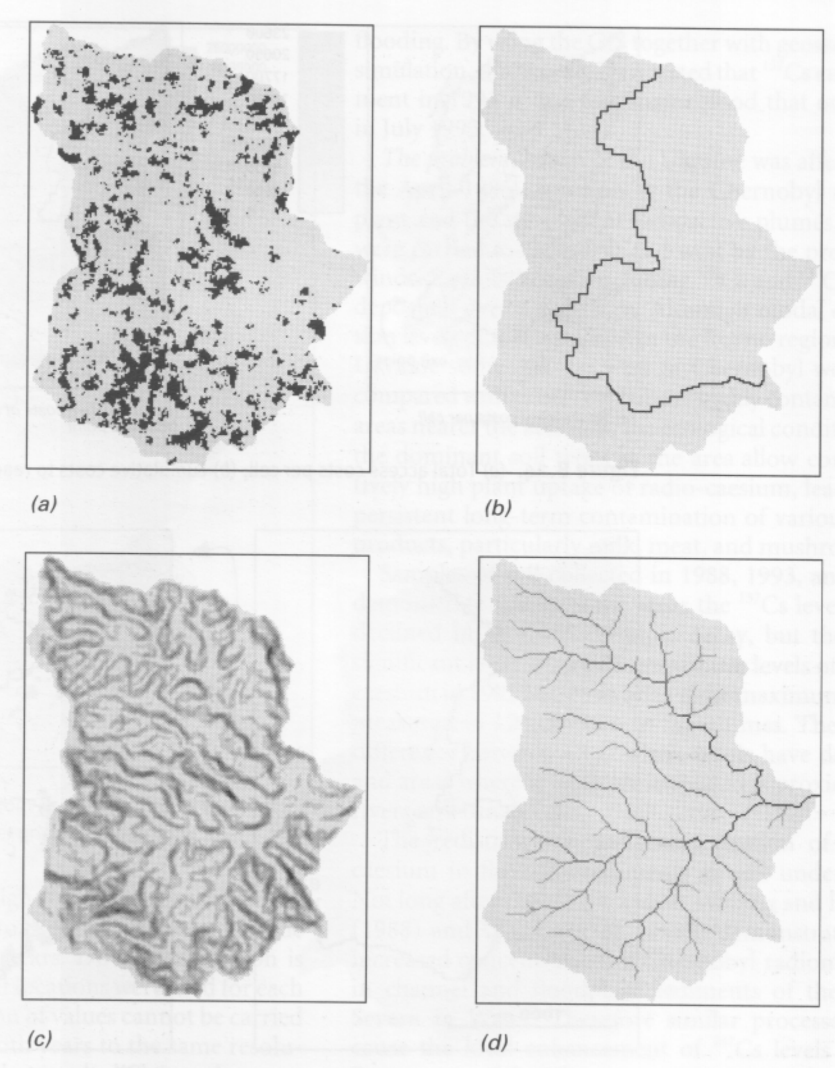


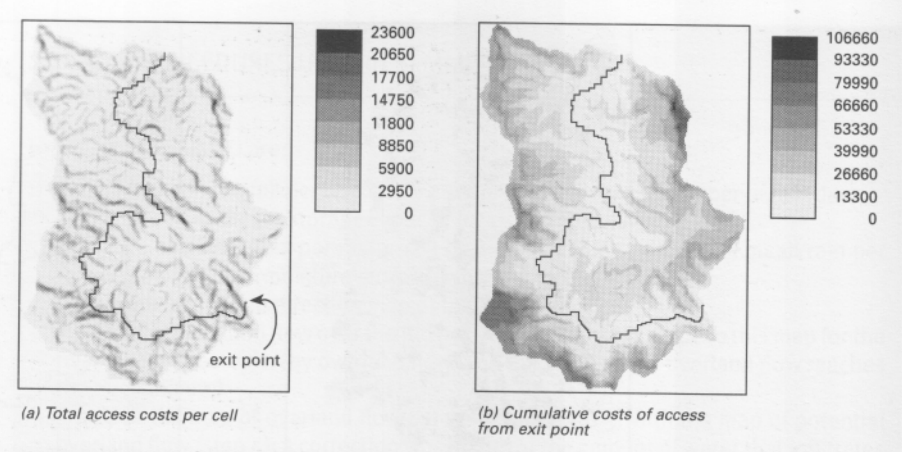
Figure 8.23. The components of the costs of timber extraction (*a*) tree locations, (*b*) access road, (*c*) slopes, (*d*) streams

Figure 8.24*a* shows the total costs per cell for the whole area.

3. The *cumulative costs of access* are computed by a spreading operation starting at the entry to the forest using the cost of cell traverse as a friction. The full set of optimal paths over the resulting continuous surface is then derived in just the same way as a stream net is derived for a topographical DEM. Figure 8.24*b* shows the cumulative costs of access to reach all cells along the optimum paths.

4. The subset of the optimum paths needed to reach the trees is achieved by accumulating 'tree flow' from the tree locations over the cumulative costs of access.

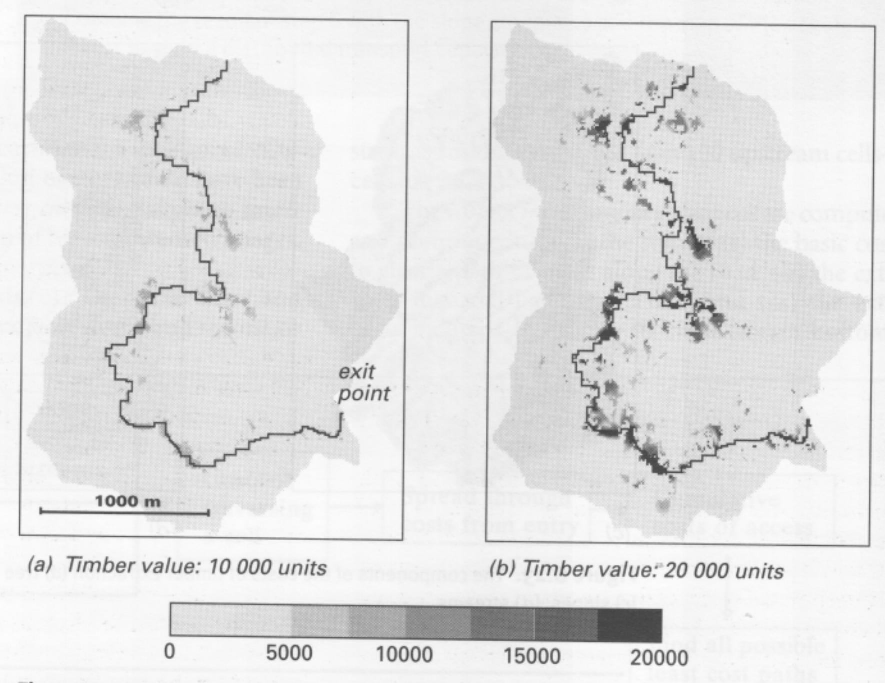
5. The cumulative increase in timber value over these optimum paths can be computed by calculating the current market price of timber by multiplying the tree locations map by the current price and using this as a friction value when spreading upstream over this subset of optimal paths. The resulting surface (Figure 8.25) gives the cumulative increase in timber value as a function of optimum cost distance from the entry point. Subtracting the costs of access from this map of *cumulative tree values* gives a map that shows the



(a) Total access costs per cell

(b) Cumulative costs of access from exit point

Figure 8.24. (*a*) Total access costs per cell; (*b*) cumulative costs to reach each cell from exit point



(a) Timber value: 10 000 units

(b) Timber value: 20 000 units

Figure 8.25. Timber stands with values > access costs

areas that can be profitably extracted (Figure 8.26). The analysis can be repeated for different market prices to see how the profitability of timber extraction varies both spatially and in gross monetary value.

Note how the optimum cost paths can be modified to take account of other cost factors. Note also that it would be easy to modify the procedure to show not just tree locations but variations in timber value over the area due to different tree species or different stages of maturity. The analysis could also be linked to a tree growth model to show how the exploitation of the forest could be conducted in a sustainable way.

The redistribution of Chernobyl ^{137}Cs by flooding

This example presents an interesting problem: data on the same attribute, in this case radiocaesium levels in soil, were collected from sets of point locations in 1988 and 1993 (Burrough *et al.* 1996). Remarkably, many

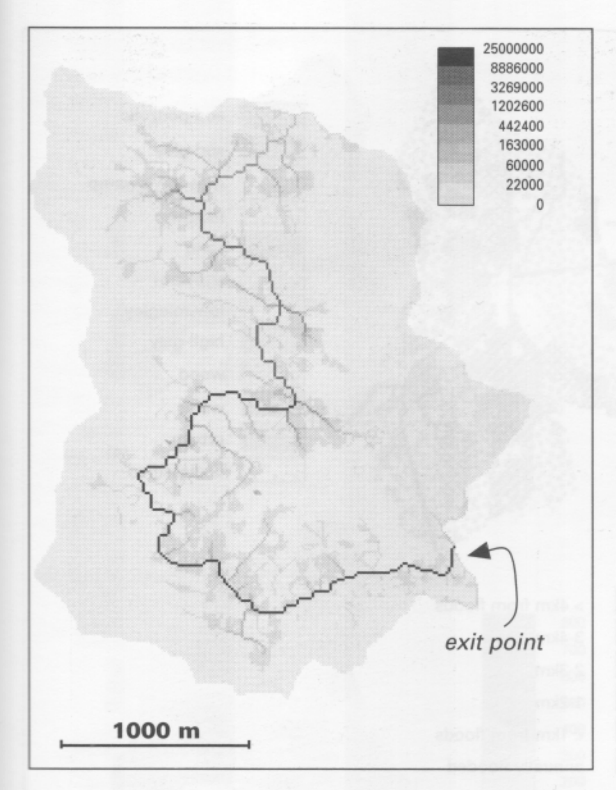


Figure 8.26. Cumulative profit and access routes to profitable stands when timber price is 20 000 units

of the 1993 levels were higher than those measured in 1988, which is contrary to expectations given that the half-life of ^{137}Cs is 30.2 years. Direct comparison is difficult because different locations were used for each year so simple subtraction of values cannot be carried out. Interpolation for both years to the same resolution grid is possible but the simple difference between the interpolated surfaces does not necessarily indicate real change in ^{137}Cs levels because the differences between the years could be caused by sampling a highly variable pattern or by a physical process, such as flooding. By using the GIS together with geostatistical simulation, the hypothesis is tested that ^{137}Cs enhancement in 1993 is due to a major flood that occurred in July 1993.

The problem. Much of the Ukraine was affected by the April 1986 explosion at the Chernobyl nuclear plant and the subsequent radioactive plumes, which were carried to the north and west by the prevailing winds. Radioisotopes, including ^{134}Cs and ^{137}Cs, were deposited over a wide area. Although modal deposition levels of radiocaesium in the Rovno region of the Ukraine some 350 km west of Chernobyl were low compared with other much more highly contaminated areas nearer the accident, the ecological conditions of the dominant soil types in the area allow comparatively high plant uptake of radio-caesium, leading to persistent long-term contamination of various food products, particularly milk, meat, and mushrooms.

Samples of soil collected in 1988, 1993, and 1994 demonstrate that in some areas the ^{137}Cs levels have declined in line with isotopic decay, but there are significant areas where the maximum levels of radiocaesium in 1993 and 1994 exceed the maximum levels measured in 1988 by two to three times. The major difference between areas where levels have declined and areas where levels have increased is proximity to rivers and canals.

The redistribution and concentration of radiocaesium in fluviatile sediments is well understood. Not long after the 1986 accident Walling and Bradley (1988) and Walling *et al.* (1989) demonstrated the increased concentration of Chernobyl radionuclides in channel and floodplain sediments of the River Severn in Wales. Therefore similar processes may cause the local enhancement of ^{137}Cs levels in the Rovno area of the Ukraine.

The data were collected from a typical collective farm area within the Rovno area (the Chapayev and Kolos collective farms) which had been identified in 1988 by the Ukrainian authorities as a site having

Table 8.2. Statistics of the transformed data as 1986 equivalents (kBqm^{-2})

Year	N	Min	Max	Mean	Median	Mode	Cv
1988	72	9.3	406.0	134.4	97.6	93.0	67.1
1993	87	20.6	1267.8	239.4	164.6	107.3	95.2
1994	47	17.3	531.5	135.1	135.1	multiple	83.0

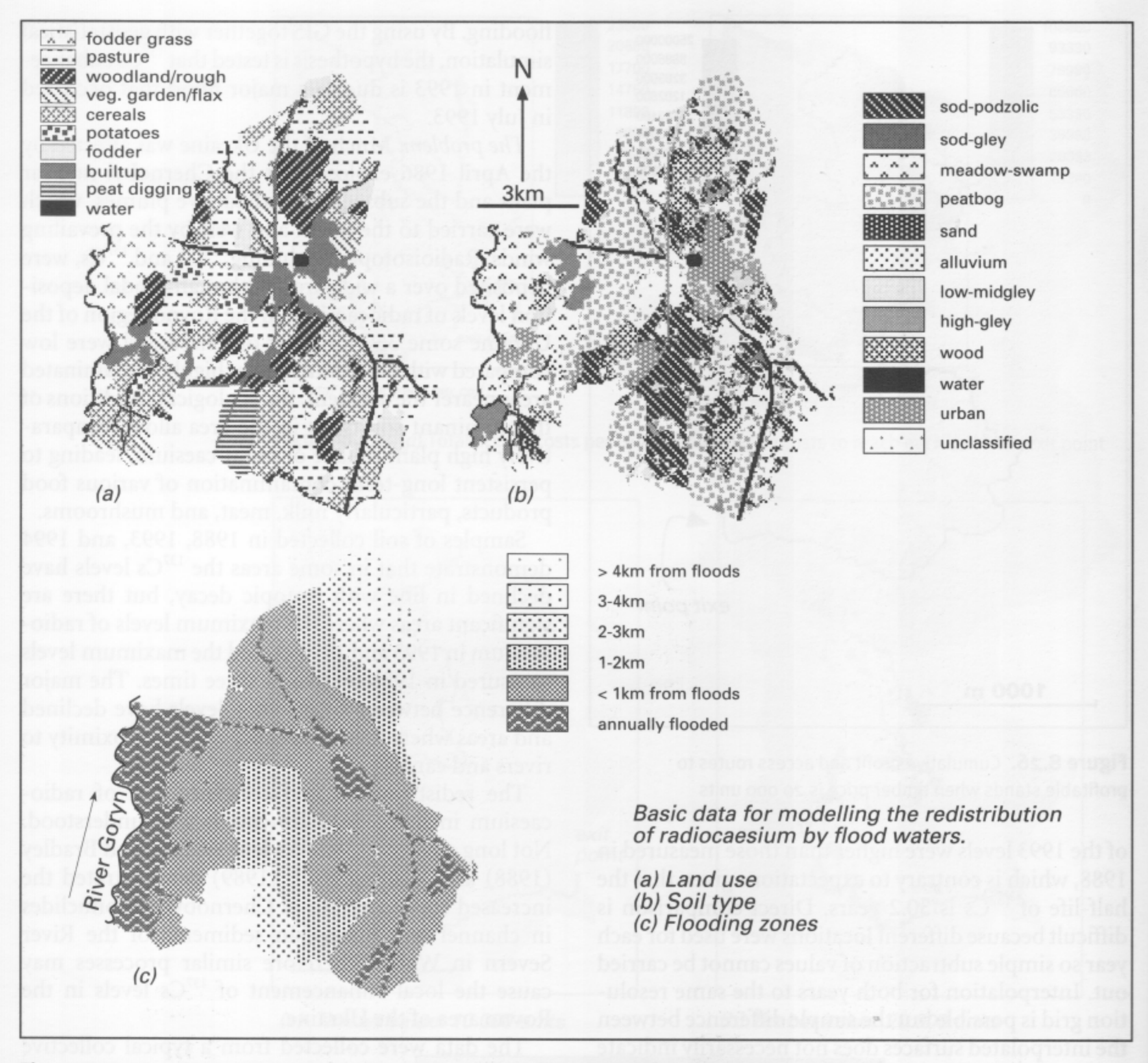


Figure 8.27. Input maps for modelling the redistribution of ^{137}Cs by flooding

a large background level of radiocaesium with strong spatial and temporal variability. The topography is extremely flat and the main drainage is by meandering river channels that drain an area of approximately 6500 km^2 in a northerly direction, uniting north of the Ukraine–Belarus border to flow east as the River Pripyat which joins the Dnieper River at Chernobyl. The whole Belarus–Ukraine border area is characterized by extensive areas of soddy gley and peaty soils which are marked on modern and historical 1 : 3 000 000 and 1 : 500 000 maps. The actual study site measures some 10.5 × 12.5 km and covers 76.5 km^2 of the Dubrovitsa District of the Rovno region.

A database of sampling locations, natural and artificial drainage, areas affected by annual spring floods, field parcel numbers, soil series, and land use maps was created by digitizing 1 : 10 000 scale paper source maps (Figure 8.27*a*,*b*). A gridded database with a resolution of 50 × 50 m was created from the basic map data for quantitative spatial analysis and interpolation. Samples of contaminated soil were collected in 1988, 1993, and 1994 at 72, 87, and 47 sites respectively and were analysed for ^{134}Cs and ^{137}Cs. The location of each sample was digitized (Figure 8.28*a*,*b*).

Correction for ^{134}Cs and decay. To study temporal changes in the levels of ^{137}Cs, which had possibly been

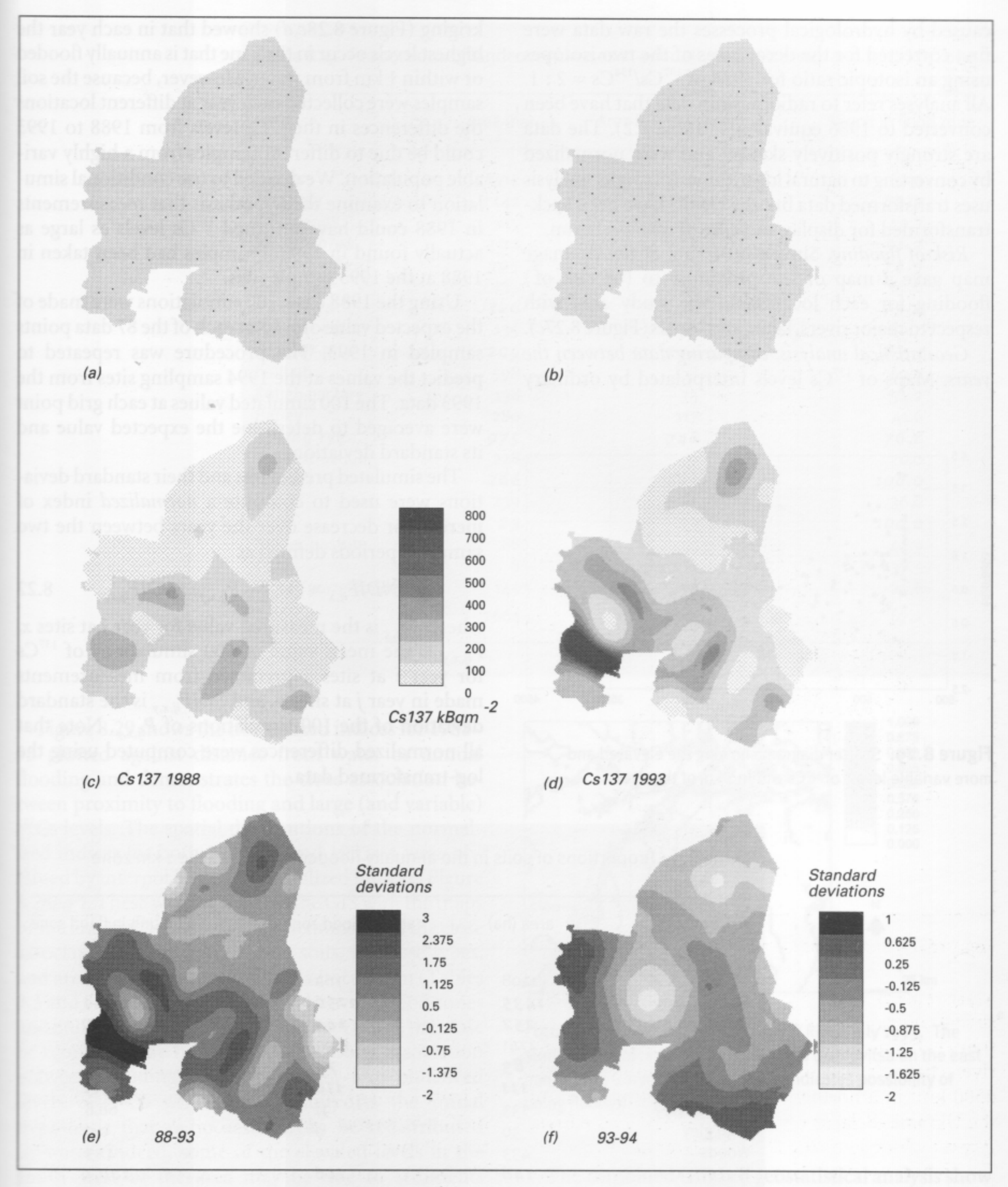


Figure 8.28. (*a,b*) Location of data points in 1988 and 1993; (*c,d*) Interpolation of ^{137}Cs levels for 1988 and 1993; (*e,f*) Normalized differences computed for 1988–93 and 1993–94, respectively

caused by hydrological processes the raw data were first corrected for the decay rates of the two isotopes using an isotopic ratio for 1986 of $^{137}Cs/^{134}Cs = 2 : 1$. All analyses refer to radiocaesium data that have been converted to 1986 equivalents (Table 8.2). The data are strongly positively skewed and were normalized by converting to natural logarithms. All spatial analysis uses transformed data but the results have been back-transformed for display and ease of interpretation.

Risk of flooding. Simple buffering of the drainage map gave a map of site proximity to (or risk of) flooding for each location in the study area with respect to major rivers, lakes, and canals (Figure 8.27*c*).

Geostatistical analysis: comparing data between the years. Maps of ^{137}Cs levels interpolated by ordinary kriging (Figure 8.28*c*,*d*) showed that in each year the highest levels occur in the zone that is annually flooded or within 1 km from water. However, because the soil samples were collected each year at different locations the differences in the ^{137}Cs levels from 1988 to 1993 could be due to different samples from a highly variable population. We decided to use conditional simulation to examine the hypothesis that measurements in 1988 could have returned ^{137}Cs levels as large as actually found in 1993 if samples had been taken in 1988 at the 1993 sample sites.

Using the 1988 data, 100 simulations were made of the expected value of ^{137}Cs at each of the 87 data points sampled in 1993. The procedure was repeated to predict the values at the 1994 sampling sites from the 1993 data. The 100 simulated values at each grid point were averaged to determine the expected value and its standard deviation.

The simulated predictions and their standard deviations were used to compute a *normalized* index of increase or decrease over the years between the two sampling periods defined as

$$NDIF_{ij,x} = (M_{i,x} - P_{ij,x,y}) \,/\, SDIF_{ij,x,y} \qquad 8.22$$

where: $M_{i,x}$ is the measured value for year i at sites x, $P_{ij,x,y}$ is the mean value of 100 simulations of ^{137}Cs for year i at sites x predicted from measurements made in year j at sites y, and $SDIF_{ij,x,y}$ is the standard deviation of the 100 simulations of $P_{ij,x,y}$. Note that all normalized differences were computed using the log-transformed data.

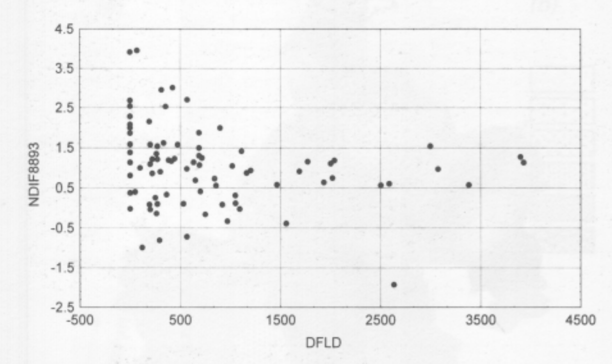


Figure 8.29. Scatter diagram showing the elevated and more variable levels of ^{137}Cs within 1 km of the rivers

Table 8.3. Proportions of soils in the annually flooded areas and the 1 km zone

Soil group	area (ha)	area in flood zone (ha)	% unit in flood zone
Soddy podzol	1408	762	54.1
Soddy gley	***1435***	***1250***	***87.1***
Meadow-swamp	***252***	***244***	***96.0***
Peatbog	2741	1605	58.5
Sand	***83***	***74***	***89.1***
Alluvium	***111***	***110***	***99.1***
Low/midgley	133	91	68.8
High gley	30	0	0.0
Woods	433	179	41.3
Water	***146***	***146***	***100.0***
Urban	503	366	72.7
Unclassified	***393***	***314***	***79.9***
Total	7668	5140	67.0)

Table 8.4. Proportions of landcover classes in annually flooded areas and the 1 km zone

Landuse	Area (ha)	Area in flood zone (ha)	% unit in flood zone
Fodder	1412	932	66.0
Pasture	***1141***	***1018***	***89.2***
Woodlands	915	451	49.3
Vegetables	***302***	***276***	***91.2***
Spring wheat	***305***	***229***	***74.9***
Maize	133	88	66.0
Potatoes	***126***	***92***	***73.2***
Fodder beet	135	82	60.7
Fodder mix	***209***	***174***	***83.3***
built up	598	418	69.9
Rough pasture	138	75	54.2
Barley	280	117	41.6
Spring rye	***975***	***748***	***76.8***
Peat extraction	228	0	0.0
Water	***266***	***266***	***100.0***
Rye	239	88	36.8
Flax	***23***	***23***	***100.0***
fodder-pasture	88	40	44.9
Oats	8	3	37.5
Wood/graze/fodder	145	34	23.6
Total	7664	5154	67.2

Figure 8.29 shows the normalized indices for 1988–93 plotted against distance from water or annual flooding and demonstrates the close association between proximity to flooding and large (and variable) ^{137}Cs levels. The spatial distributions of the normalized indices for both 1988–93 and 1993–4 were obtained by interpolating the normalized indices (Figure 8.28*e,f*). These were overlaid in the GIS with the maps of soil and land use; these maps demonstrated the close association between floodplain soils, land use types, and areas with elevated levels of radiocaesium (Tables 8.3 and 8.4). Although the match between flood zones and enhanced normalized levels is not exact, the close fit reinforces the hypothesis of a strong association between flooding, flood potential, and enhanced levels of ^{137}Cs which clearly supports the initial hypothesis that radiocaesium can be redistributed by water: indeed, some of the elevated levels in the south–west of the area may be due to sediments being transported from further upstream (cf. Walling and Bradley 1988, Walling *et al.* 1989).

Entries in bold type denote classes with a larger proportion within the flooded area than in the area as a whole.

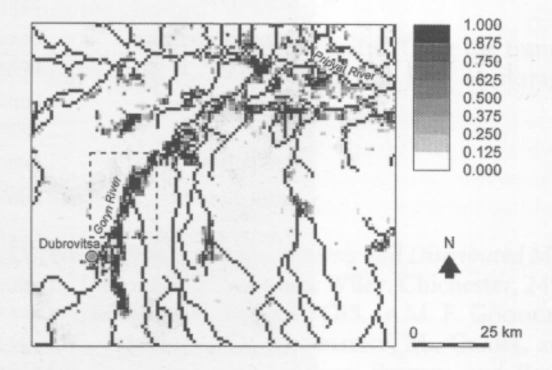


Figure 8.30. Flooding in the Pripyat River, July 1993. The study area is located just north-east of Dubrovitsa on the east bank of the Goryn River. Grey scale indicates possibility of being flooded

The combined GIS and geostatistical analysis show clearly how data on flood frequency, soil types, land use, and soil samples can be combined to evaluate the hypothesis that the enhanced levels of ^{137}Cs measured after 1988 are due to flooding and deposition of polluted sediments. Further data on the enhancement

of ^{137}Cs levels in milk before and after a major flood in July 1993 confirmed the direct effect of flooding on uptake into the foodchain, particularly on unimproved pastures on peaty soils. After the study was nearly completed we found a set of NOAA images on the Internet which confirmed the flood of July 1993 and provided more support for the movement of ^{137}Cs by floodwaters in the Ukraine (Figure 8.30).

Conclusions

This chapter has shown a wide range of generic methods for deriving new data from attributes modelled as continuous surfaces. Table 8.5 summarizes the functional capabilities of a GIS that can deal both with exact entities and continuous fields. Frequently, but not always do the continuous surfaces represent landform. These methods have applications in many fields, not just including hydrology, but also erosion and land

Table 8.5. Functional capabilities of GIS for analysis of entities and continuous fields

Geometric	Conversion of geographic coordinates from one correction cartographic projection to another
Interactive	Interactive updating and editing of geographic editing and attribute data
Sorting	Sort attribute or geographic data as required
Location	Locate entities having defined sets of attributes
Summarize	Summarize attributes per geographic entity (point/line/polygon/cell)
Compute statistics	Compute statistics (means, areas, enclosures, etc.) for points/lines/polygons/cells
Proximity	Conduct nearest neighbour and proximity searches—create buffer zones and carry out corridor analyses
Interpolate	Interpolate from point data to regular grid or isolines (contours)
Block diagram	Compute block diagrams of three-dimensional data
Overlay analysis	Overlay and combine several maps in either vector (polygon) or raster (grid) mode using Boolean (AND/OR/NOT) logic and arithmetical functions/filters to manipulate both the geographic and the attribute data
Polygon to raster	Convert graphic representation from polygon to grid cell representation
Edge detection	Semi-automated detection of edges of images in raster representation
Network analysis	Find shortest path along a road network with weighted route parameters for traffic density. Link specific entities (roads, cables) to each other and follow the network through
Digital terrain analysis	Represent landform as a 3D surface. Compute slope, aspect, intervisibility, rate of change of slope, shaded relief, direction of flow, determine watershed boundaries.
Models	Ability to interface with simulation models

degradation studies, forest management and soil and water pollution. Although all the examples relate to physical problems of the landscape, these kinds of analyses can also be applied to any study in which one or more attributes can be modelled as a continuous surface, such as surfaces of market potential, exposure to disease, or economic well-being.

Questions

1. Compare and contrast the effects of (a) spatial filtering and (b) polygon reclassification and entity merging, for generalizing a soil or land use map.

2. Explain why digital elevation models are essential prerequisites for modelling environmental processes. Derive a standard procedure for estimating the range of ecological sites in an area from gridded DEMs before commencing fieldwork.

3. Explore the assumptions and difficulties in collecting data for a regional erosion model such as SEMMED.

4. Devise a suitable set of spatial analysis operations for deriving the best location of hiking trails in a national park, taking into account the wish to have good views but to avoid difficult or dangerous terrain.

5. Explain how you would use the tools presented in this chapter to investigate how the following organisms or societies might exploit a given part of the earth's surface:
   (*a*) Asterix's village in Brittany.
   (*b*) Black mould growing on the bathroom wall.
   (*c*) Owners of fast food outlets in a city.
   (*d*) Polynesian seafarers in the Pacific Ocean.
   (*e*) The dispersion of plants and animals along hedgerows.
   (*f*) Manufacturers of solar heating systems.

6. Discuss the physical factors affecting the transport capacity *TC* for the transport of sediment over a landscape and explain how they affect the flow or storage of sediment.

Suggestions for further reading

Beven, K. J., and Moore, I. D. (eds.) (1994). *Terrain Analysis and Distributed Modelling in Hydrology*. Advances in Hydrological Processes, Wiley, Chichester, 249 pp.

Maidment, D. R. (1996). Environmental modeling with GIS. In M. F. Goodchild, L. T. Steyaert, B. O. Parks, C. Johnston, D., Maidment, M., Crane, and S. Glendinning (eds.), *GIS and Environmental Modeling: Progress and Research Issues.* GIS World Books, Fort Collins, Colo., pp. 315–24.

Moore, I. D. (1996). Hydrological modeling and GIS. In M. F. Goodchild, *et al., GIS and Environmental Modeling: Progress and Research Issues,* 143–8.

Tomlin, C. D. (1990). *Geographic Information Systems and Cartographic Modeling.* Prentice Hall, Englewood Cliffs, NJ, 249 pp.

NINE

Errors and Quality Control

This chapter examines how errors occur in spatial data and the effects that they may have on data analysis and modelling. Errors include blunders and gaffs, but they are also an intrinsic part of the choice of data models and computational models. Statistical uncertainty and spatial variation are critical aspects of any error analysis in spatial data. Methods are presented for estimating errors in the entity domain in vector-raster conversion, digitizing, and polygon overlay.

Spatial data, costs, and the quality of GIS output

The quality of GIS products is often judged by the visual appearance of the end-product on the computer screen, plotter, or video device, and computer cartographers are devising ever more appealing techniques for communicating visual information to people. Quality control by visual appearance is insufficient, however, if the information presented is wrong or is corrupted by errors. Uncertainties and errors are intrinsic to spatial data and need to be addressed properly, not swept away under the carpet of fancy graphics displays. There can be a false lure about the attractive, high-quality cartographic products that cartographers, and now computer graphics specialists, provide for the users of GIS. In the 1980s Chrisman (1984*a*) pointed out that 'we have developed expectations, such as smooth contour lines, which are not always supported by adequate evidence' and Blakemore (1984) drew attention to the naïve claims of some adherents of computer cartography that computer-assisted cartographic products are necessarily accurate to the resolution of the hardware used to make them. He noted that only a few critical authors, such as Boyle (1982), Goodchild (1978), Jenks (1981), and Poiker (1982), had drawn attention to the problems of errors in geographic information processing but in 1996 even after twenty-five years of development there is still inadequate attention to how errors arise and are propagated. Most studies on errors are still at the research level (Fisher 1995, Goodchild and Gopal 1989, Heuvelink 1993, Lodwick *et al.* 1990) though systematic studies of spatial data quality are now being published (Guptill and Morrison 1995).

In a recent study, Wellar and Wilson (1995) conclude that though GIS has had an impact on the qualitative, quantitative, and/or visualization procedures of spatial theorizing, it has had little impact on the process of spatial theorizing, and hence on a better understanding of natural variation and errors. This is surprising given the costs of data acquisition and the investments that are linked to the use of GIS. In the fields of geostatistics and spatial statistics, however, there have been many theoretical and practical studies on how to deal with uncertainty in the spatial variation of attributes that can be treated as continu-

ous fields (e.g. Isaaks and Srivastava 1989, Journel 1996, Deutsch and Journel 1992, Cressie 1991) and it is time to link the ideas developed in these areas to provide a sound basis for understanding the role of uncertainty in spatial data and spatial data analysis.

Data accuracy is often grouped according to *thematic accuracy*, *positional accuracy*, and *temporal accuracy* (Aalders 1996) but errors in spatial data can occur at various stages in the process from observation to presentation. Errors in perception (improper identification) can occur at the conceptual stage. Errors and approximations in determining the geographical location depend on surveying skills, the provision of hardware (GPS satellites, laser theodolites, etc.) and the choice of map projections and spheroids. Errors in the measurement of attributes, due to variation in the phenomenon in question, the accuracy of the measurement device, or observer bias can occur during the recording of the primary data. For phenomena treated as continuous fields, the density of samples, their support (physical size of the sample), and the completeness of the sampling are all sources of uncertainty.

Errors can creep in when data are stored in the computer because too little computer space is allocated to store the high-precision numbers needed to record data to a given level of accuracy. Some data may be so expensive or difficult to collect that one must make do with a few samples and rely on inexact correlations with other, cheaper to measure attributes, and so inevitably uncertainties arise. The logical or mathematical rules ('models' or interpolations) used to derive new attributes from existing data may be flawed or may involve computational methods that lead to rounding errors. When data that have been measured on different entities, or sampled on different supports are combined, the differences in spatial resolution may be so great that simple comparisons cannot be made.

Finally, in the visual presentation of results, users can obtain erroneous impressions if the semiotic language is not clear, if colours and shading are inappropriate or if displays are too crowded, or if it is just too difficult to get a clear result.

The usual view of errors and uncertainties is that they are bad. This is not necessarily so, however, because it can be very useful to know how errors and uncertainties occur, how they can be managed and possibly reduced, and how knowledge of errors and error propagation can be used to improve our understanding of spatial patterns and processes. Linking a good understanding of spatial uncertainties to numerical methods of modelling and interpolation can provide useful tools for optimizing sampling (and thereby improving value for money) and for identifying the weak and strong parts of spatial analysis. A good understanding of errors and error propagation leads to active quality control.

Many GIS users conduct data analysis using the techniques presented in Chapters 7 and 8 under the implicit assumption that all data are totally error free. By 'error free' is meant not only the absence of factually wrong data caused by faulty survey or input, but also statistical error, meaning free from variation. In other words, the arithmetical operation of adding two maps together by means of a simple overlay implies that both source maps can be treated as perfect, completely deterministic documents with uniform levels of data quality over the whole study area. This view is imposed to a large extent by the absence of information about data quality, the exact concepts embodied in most databases and retrieval languages (though it can be otherwise), a lack of understanding of how errors are propagated, and the absence of GIS tools for error evaluation.

Many field scientists and geographers know from experience that carefully drawn boundaries and contour lines on maps are elegant misrepresentations of changes that are often gradual, vague, or fuzzy (Burrough and Frank 1996). People have been so conditioned to seeing the variation of the earth's surface portrayed either by the stepped functions of choropleth maps (sharp boundaries) or by smoothly varying mathematical surfaces (see Chapter 2) that they find it difficult to conceive that reality is otherwise. Besides the 'structure' that has been modelled by the boundaries or the isolines, there is very often a residual unmapped variation that occurs over distances smaller than those resolvable by the original survey. Moreover, the spatial variation of natural phenomena is not just a local noise function or inaccuracy that can be removed by collecting more data or by increasing the precision of measurement, but is often a fundamental aspect of nature that occurs at all scales, as the proponents of fractals have pointed out (see Mandelbrot 1982, Burrough 1983*a*,*b*, 1984, 1985, 1993*a*, Goodchild 1980, Lam and De Cola 1993).

It is very important to understand the nature of errors in spatial data and the effect they may have on the quality of the analyses made with GIS.

The first part of this chapter explores the sources of errors in spatial data, and the factors affecting their quality, both with respect to entity-based and continuous field-based models of spatial phenomena. The second examines the factors that affect the quality of spatial data, while the third covers the development and understanding of errors associated with transforming entity and field-based data from one representation to another (vector-raster), by line digitizing, and through polygon overlay. Chapter 10 presents a statistical approach to the understanding of error propagation in numerical modelling in the context of the kinds of spatial analysis presented in Chapters 6 and 7 and shows how a proper understanding of uncertainties can be used for optimizing sampling and spatial analysis.

Sources of errors in spatial data

Box 9.1 shows the main factors governing the errors that may be associated with geographic information processing. The word 'error' is used here in its widest sense to include not only 'faults' but also to include the statistical concept of error meaning 'variation'. The 'errors' include faults that are obvious and easy to check on but there are more subtle sources of error that can often only be detected while working intimately with the data. The most difficult sources of 'errors' are those that can arise as a result of carrying out certain kinds of processing; their detection requires an intimate knowledge of not only the data, but also the data models, the data structures, and the algorithms used. Consequently they are likely to evade most users. Many of these aspects of 'error', or more correctly 'data quality', are being addressed through international agreements (cf. Aalders 1996).

Accuracy of content

The accuracy of content is the problem of whether the attributes attached to the points, lines, and areas of the geographic database are correct or free from bias. We can distinguish between qualitative accuracy, which refers to whether nominal variables or labels are correct (for example, an area on a land use map might be wrongly coded as 'wheat' instead of 'potatoes') and quantitative accuracy which refers to the level of bias in estimating the values assigned (for example a badly calibrated pH meter might consistently estimate all pH values 1 unit high). Ensuring accuracy is a matter of having reliable, documented input and transformation procedures.

Measurement errors

Poor data can result from unreliable, inaccurate, or biased observers or apparatus. The reader should clearly understand the distinction between accuracy and precision. Accuracy is the extent to which an estimated value approaches the true value and is usually estimated by the standard error. In statistical terminology, precision is a measure of the dispersion (usually measured in terms of the standard deviation) of observations about a mean. Precision also refers to the ability of a computer to represent numbers to a certain number of decimal digits.

Field data

The surveyor is a critical factor in the quality of data that are put in to many geographical information systems. Well-designed data collection procedures and standards help reduce observer bias. The human factor is most important in data collection methods relying on intuition such as in soil or geological survey where an interpretation is made in the field, or from aerial photographs or seismographs, of the patterns of variation in the landscape or subsurface. The user should realize that some observers are inherently more perceptive or industrious than others—'the quality of soil surveyors varies from the two minute job of an irresponsible aerial photo interpreter to that of the surveyor whose sampling plan suggests that he is planting onions' (Smyth, quoted in Burrough 1969). Very large differences in the appearance of a map can result from differences in surveyor or from mapping methods as studies by Bie and Beckett (1973) and

BOX 9.1. FACTORS AFFECTING THE QUALITY OF SPATIAL DATA

1. Currency
   - Are data up to date?
   - Time series
2. Completeness
   - Areal coverage—is it partial or complete?
3. Consistency
   - Map scale
   - Standard descriptions
   - Relevance
4. Accessibility
   - Format
   - Copyright
   - Cost
5. Accuracy and Precision.
   - Density of observations
   - Positional accuracy
   - Attribute accuracy—qualitative and quantitative
   - Topological accuracy
   - Lineage—When collected, by whom, how?
6. Sources of errors in data
   - Data entry or output faults
   - Choice of the original data model
   - Natural variation and uncertainty in boundary location and topology
   - Observer bias
   - Processing
     - Numerical errors in the computer
     - Limitations of computer representations of numbers
7. Sources of errors in derived data and in the results of modelling and analysis
   - Problems associated with map overlay
   - Classification and generalization problems
   - Choice of analysis model
   - Misuse of logic
   - Error propagation
   - Method used for interpolation

more recently Legros *et al.* (1996) for soil survey and Salome *et al.* (1982) for geomorphology have clearly demonstrated.

In large survey organizations it should be possible to determine and record the qualities of each surveyor, an extra attribute that could be stored with the data themselves. Such procedures might be resisted by the staff as a slur on professional expertise but the best method for improving observer quality is to improve all aspects of the data-gathering process, such as standardizing observational techniques and data recording forms and by developing a joint commitment between survey management and staff to work to the highest possible standards.

The increasing use in many field sciences of automated sampling devices linked to electronic data loggers means that if all is operating properly then the accuracy and the precision of the data are good. Data from electronic sampling devices collectors can be automatically screened for extreme values indicating

malfunctioning. New sampling devices in areas such as geoengineering and pollution science mean that observations can be made in situ of materials that otherwise must be analysed in the laboratory (Rengers 1994).

Laboratory errors

Intuitively, one expects the quality of laboratory determinations to exceed those made in the field. Although determinations carried out within a single laboratory using the same procedure may be reproducible, the same cannot be said of analyses performed in different laboratories. The results of a major worldwide laboratory exchange program carried out by the International Soil Reference and Information Centre in Wageningen (van Reeuwijk 1982, 1984) showed that variation in laboratory results for the same soil samples could easily exceed ±11 per cent for clay content, ±20 per cent for cation exchange capacity (±25 per cent for the clay fraction only), ±10 per cent for base saturation, and ±0.2 units for pH. The implications for the results of numerical modelling are enormous! Laboratory analyses should be improving in reproducibility thanks to the wider use of automated laboratory equipment, but no amount of laboratory technology will make up for poorly collected or poorly prepared samples.

Locational accuracy

The importance of the locational accuracy of geographic data depends largely on the type of data under consideration. Topographical data are usually surveyed to a very high degree of positional accuracy that is appropriate for the well-defined objects such as roads, houses, land parcel boundaries, and other features that they record. With modern techniques of electronic surveying and GPS the position of an object on the earth's surface can now be recorded to millimetre accuracy. In contrast, the position of soil or vegetation unit boundaries often reflects the judgement of the surveyor about where a dividing line, if any, should be placed. Very often, vegetation types grade into one another over a considerable distance as a result of transitions determined by microclimate, relief, soil, and water regimes. Changes in slope class or groundwater regime are also unlikely to occur always at sharply defined boundaries.

Positional errors can result from poor fieldwork, through distortion or shrinkage of the original paper base map or through poor quality vectorizing after raster scanning (Dunn *et al.* 1990, Bolstad *et al.* 1990). Local errors can often be corrected by interactive digitizing on a graphics work station, while general positional errors can be corrected by various kinds of transformation, generally known as 'rubber-sheeting' techniques, that have been described in Chapter 4. The combination of modern hardware and software for error detection has greatly improved the quality of digitizing in recent years.

The success of rubber-sheeting methods for correcting geometrical distortion depends largely on the type of data being transformed, and the complexity of the transformations. Many methods work well for simple linear transformations but break down when complex shrinkages must be corrected. The methods do not necessarily work well when the original map consists largely of linked, straightline segments. For example, some years ago, attempts were made at the Netherlands Soil Survey Institute to use rubber-sheeting methods to match a digitized version of an early nineteenth-century topographic map on to a modern 1 : 25 000 topographical sheet for the purpose of assessing changes in land use. The road pattern of the area in question was similar to that of a rigid girder structure. When submitted to the rubber-sheeting process, the road lines were not stretched but the structure crumpled at the road intersections, in much the same way that a bridge or crane made from meccano might crumple at the joins!

Natural spatial variation

Many thematic maps, particularly those of natural properties of the landscape such as soil or vegetation, do not take into account local sources of spatial variation or 'impurities' that result from short-range changes in the phenomena mapped. This problem has been the subject of much research, particularly in soil survey, soil physics, and groundwater studies (e.g. Beckett and Webster 1971, Bouma and Bell 1983, Nielsen and Bouma 1985, Burrough 1993*b*). The problems are as much associated with paradigms of soil classification and mapping that are spatially simplistic as with the natural variation of the soil which was incompletely understood (Burrough *et al.* 1997).

Cartographic conventions forced soil scientists to map soils as crisply delineated, homogeneous areas. Information about gradual change within boundaries, and boundaries of varying width could not be represented on conventional chorochromatic maps. These maps have been diligently digitized and the digital soil polygon has been presented to GIS users as an entity

that is spatially as well-defined as a cadastral unit. Unfortunately, the truth is often otherwise—these crisp polygons are really crude, but convenient approximations, and a major problem concerns the lack of information about the difference between these models and reality.

Initially, conventional soil series maps at scales of 1 : 25 000–1 : 50 000 were characterized in terms of the 'impurities' within the units delineated (Soil Survey Staff 1951), which were supposed to be no more than 15–25 per cent. Impurities were defined as observations that did not match the full requirements as specified in the map legend. Many studies (e.g. see Beckett and Webster 1971 or Burrough 1993*b* for a review) have shown that not only was the 15 per cent a wild guess, but that the concept of 'impurity' had little meaning. By varying the legend, the definition of just what was a matching observation, and so the purity, could be manipulated at will. Subsequent work has demonstrated the natural variation of soil and shown its importance for understanding pollution problems or for optimizing soil fertilization in precision agriculture (e.g. Burrough 1993*b*, Goode 1997). There is increasing information on the variability of soil, and other natural phenomena such as water quality or species composition, which as yet may only be available to specialists.

It is important to realize that the unseen spatial variation of phenomena like soil, lithology, or water quality can contribute greatly to the relative and absolute errors of the results produced by modelling and map overlay. More details about how to estimate how these errors propagate through numerical models are given below in this chapter.

Factors affecting the reliability of spatial data

Age of data

It is rare that all data are collected at the same time for a given project, unless that project is a specific piece of research. Most planners and environmental agencies are forced to use existing published data in the form of maps and reports, filled in as necessary by more recent remote sensing imagery (including aerial photographs) and field studies. Mead (1982) comments that 'with the exception of geological data, the reliability of data decreases with age'. Although this may be broadly true in the sense that geology changes much more slowly than soil, water regimes, vegetation, or land use, it is also possible that old data are unsuitable because they were collected according to systems of standards that are no longer used or acceptable today. Many attempts to capture old data, using handwritten field sheets and out-of-date terminology have had to be abandoned simply because of the enormous costs involved.

Areal coverage

It is desirable that the whole of a study area, be it an experimental field or a country should have a uniform cover of information. If this is not so the resource data processer must make do with partial levels of information. Though global digital data are increasing in availability (e.g. the Digital Chart of the World on Internet) it is still common, even in developed countries, for there to be no complete cover of certain kinds of thematic information over a study area, except at scales that are too small for the purpose required. For example, many countries still have fragmentary coverage of soil maps at scales of 1 : 25 000–1 : 50 000. Moreover, during the 30–40 years the concepts and definitions of thematic classes of soil, vegetation, and geology have changed as have the ways they should be mapped and the surveyors themselves have moved on. Historical facts can lead to inconsequential map units along map sheet boundaries that are difficult to resolve without further survey.

If coverage is not complete, decisions must be made about how the necessary uniformity is to be achieved. Options are to collect more data, to obtain surrogate data from remote sensing, or to generalize detailed data to match less detailed data. Note that it is extremely unwise to 'blow up' generalized or small-scale map data to obtain the necessary coverage.

Map scale and resolution

Most geographic resource data have been generated and stored in the form of thematic maps, and only in

recent years with the development of digital information systems has it been possible to have the original field observations available for further processing. Large-scale maps not only show more topological detail (spatial resolution) but usually have more detailed legends (e.g. a soil map of scale 1 : 25 000 and larger usually depicts soil series legend units, while a soil map of scale 1 : 250 000 will only display soil associations-—see Vink 1963 for details). It is important that the scale of the source maps matches that required for the study—small-scale maps could have insufficient detail and large-scale maps may contain too much information that becomes a burden through the sheer volume of data. Many survey organizations provide their mapped information at a range of scales and the user should choose that which is most appropriate to the task in hand.

Density of observations

Much has been written about the density of observations needed to support a map or interpolation (e.g. Vink 1963, Burrough 1993*b*, Webster and Burgess 1984), yet there are still organizations that produce maps without giving any information whatsoever about the amount of ground truth upon which it is based. This attitude is changing—the Netherlands Winand Staring Institute provides its contract survey clients with maps showing the location and classification of all soil observations; the UK Land Resources Development Centre has published maps showing the density and location of sample points and transects in surveys (see for example the Reconnaissance Soil Survey of Sabah, Acres *et al.* 1976).

Although the actual density of observations may be a reasonable general guide to the degree of reliability of the data, it is not an absolute measure, as statistical studies of soil variation have shown. A rough guide to the density of observations needed to resolve a given pattern is given by the 'sampling theorem' originating from electronic signal detection, that specifies that at least two observations per signal element need to be made in order to identify it uniquely. There has also been considerable work in photogrammetry to estimate the densities of observations that need to be made from aerial photographs on a stereoplotter in order to support reliable digital elevation models (Makarovic 1975, Ayeni 1982).

In short, sampling density is only a rough guide to data quality. It is also important to know whether the sampling has been at an optimum density to be able to resolve the spatial patterns of interest and this subject is treated in Chapter 6 and in the next chapter.

Relevance

Not all data used in geographical information processing are directly relevant for the purpose for which they are used, but have been chosen as surrogates because the desired data do not exist or are too expensive to collect. Prime examples are the electronic signals from remote sensors that are used to estimate land use, biomass, or moisture, or observations of soil series based on soil morphology that are used to predict soil fertility, erosion susceptibility, or moisture supply. Provided that the links between the surrogates and the desired variables have been thoroughly established then the surrogates can be a source of good information.

The calibration of surrogates is a major part of remote sensing technology. Briefly, a number of pixels on the image is selected for use as a 'training set'. The variation of reflectance of each frequency band recorded is displayed in the form of a histogram; the practice is to select a training set of pixels that return narrow, unimodal distributions. These training set pixels are calibrated by 'ground-truth' observations so that the set of pixels can be equated with a crop type, a soil unit, or any other definable phenomenon. The remaining pixels in the image are then assigned to the same set as the training set using allocation algorithms based on discriminant analysis (minimum distance in multivariate space of the original frequency bands), maximum likelihood or parallelopiped classifiers (see for example, Estes *et al.* 1983, Lillesand and Kiefer 1987).

Data format, data exchange, and interoperability

There are three kinds of data format of importance. First there is the purely technical aspects of how data can be written on magnetic media for transfer from one computer system to another. This includes aspects such as the kind of medium (digital tape, floppy disk, compact disk), the density of the written information (tape block lengths, number of tracks and the density), the type of characters used (ASCII or binary), and the lengths of records. For data lines, it is essential that the speed of transmission of the two computers is matched, but most modems ensure that this is automatic.

The second kind of format concerns the way the data are arranged, or in other words, the structure of the data themselves. Do the data refer to *entities in space*, recorded as points, lines, and areas in a relational model, as *objects* in an object orientation system, or as *discretized continuous fields* coded as rasters? If the areas are coded in raster format, what is the size of each pixel? Is the organization of these data tied to a particular computer system that makes exchange difficult without conversion? For example, many commercial GIS have their own internal data structures (see Chapter 3) that may make direct data exchange difficult. The current moves to system interoperability and the availability of data sources on the Internet are driving people to develop generally acceptable, interchangeable data structures that conform to widely accepted industrial and international standards (Schell 1995*a*).

The third kind of format concerns the locational and attribute data, their scale, projection, and classification. Scale and projection conversions can usually be accomplished quite easily by using appropriate mathematical transformations on the coordinate data (e.g. Maling 1973). Matching classifications from different sources can be very difficult, and the problem is by no means confined to the problems of classifying soil profiles but also occurs in municipal applications of GIS where different administrative divisions may have completely different ways of recording essentially similar entities like roads or services.

To summarize, data exchange often requires that data be reformatted to a lowest common denominator format that can be read by many systems easily. These formats are not necessarily the most compact nor the most efficient but are expedient. There are data formats for satellite data, there are format standards for commercial vendors, there are within-country standards (e.g. in the UK for Ordnance Survey Maps, in the Netherlands and Germany for topographic mapping) and international standards for Geographical Information are now being developed. General standards for the encoding and exchange of spatial information have been set up by standards committees of the European Union (e.g. see Comité Européan Normalisation CEN Technical Committee 287—David *et al.* 1997, Salgé 1997), by the US Federal Data Standards Committee (National Research Council 1994), and by the recently formed Open GIS Consortium (Schell 1995*a*). Note that interoperability issues are forcing people to think of the conceptual problems of exchanging data as the first step, rather than solely concentrating on technical arguments.

Accessibility

Not all data are equally accessible. Data about land resources might be freely available in one country, but the same kind of data could be a state secret in another. Besides the military aspects of data for geographic information systems (here one thinks immediately of digital terrain models) inter-bureau rivalries can also obstruct the free flow of data. Costs and format problems can also seriously hinder data accessibility. In recent years a new kind of middleman, the information broker, has sprung up to assist the seeker of data from digital archives. Details about information services can be obtained from government or international agencies (e.g. EUROGI, Euronet DIANE News, the newsletter of the Directorate General for Information Market and Innovation, Commission of the European Communities, Luxembourg). There is also much information to be found on the Internet and World Wide Web (see Appendix 2).

Costs and copyrighting

Collection and input of new data or conversion and reformatting of old data cost money. For any project, the project manager should be able to assess the costs and benefits of using existing data as compared to initiating new surveys. Digitizing costs may be especially high for inputting detailed hand-drawn maps or for linking attributes to spatial data. Scanners may offer savings for data input of contour lines and photographic images. It may be cheaper for a survey agency to contract out digitizing work to specialist service bureaux than to do the work in house using staff who can be better used for more skilled work. Similarly, if an agency only occasionally needs to perform certain kinds of data transformations or to output results to expensive devices such as laser photo plotters of high quality, it may be cheaper to make use of service bureaux.

Copyright on published maps and spatial data varies from country to country and it is best to check on the legal situation in each case when digitizing maps or using spatial data for research or commercial applications (e.g. see Burrough and Masser 1997).

Numerical errors in the computer

As well as the problems inherent in the data, indicated above, there are other sources of unseen error that can originate in the computer. The most easily forgotten, yet critical aspect of computer processing is the abil-

BOX 9.2. ERROR CREATION BY COMPUTER WORD OVERFLOW

There is a simple, and very revealing test of calculation precision that can demonstrate just how computer word length can affect the results of calculation (Gruenberger 1984). The number 1.0000001 is squared 27 times (equivalent to raising 1.0000001 to the 134 217 728'th power). The table shows the results of performing this calculation on a Personal computer with a 80486 processor using Microsoft Quick Basic with 4-byte or 8-byte precision. After 27 squarings the single precision result has acquired a cumulative error of more than 1300 per cent! Clearly, the programmer must avoid situations in which the results of a calculation depend on accuracies of representation that exceed the number of digits available for representing the numbers.

No. of squares	Single precision	Double precision	Single/double per cent difference
1	1	1.00000020000001	100.0000038418561
2	1	1.00000040000006	100.0000076837067
3	1.000001	1.00000080000028	100.0000153673913
4	1.000002	1.00000160001201	100.000030734694
5	1.000004	1.00000320004962	100.0000614690337
6	1.000008	1.00000640020163	100.00012293665
7	1.000015	1.00001280081287	100.0002458676304
8	1.000031	1.00002560326416	100.0004917125829
9	1.000061	1.00005120130820 9	100.0009833344561
10	1.000122	1.000102405237993	100.0019663060907
11	1.000244	1.000204820962818	100.00393161034
12	1.000488	1.000409683877262	100.0078565185848
13	1.000977	1.000819535595404	100.0157136544184
14	1.001955	1.0016397428294	100.0314297315305
15	1.003913	1.003282174415347	100.0628687841718
16	1.007841	1.006575121499587	100.125771907418
17	1.015744	1.013193475221909	100.2517048611989
18	1.031735	1.026561018232249	100.5040404509439
19	1.064478	1.053827524154032	101.0106167959662
20	1.133113	1.110552450664617	102.0314517808015
21	1.283945	1.233326745677186	104.1041728221037
22	1.648514	1.521094861602679	108.3767906630347
23	2.717598	2.313729577994073	117.4552872926174
24	7.385337	5.353344560084633	137.9574445444612
25	54.54321	28.65829797898773	190.3225694558652
26	2974.962	821.2980430524522	362.2268061013606
27	8850397	674530.4755217875	1312.082599848986

ity of the computer to be able to store and process data at the required level of precision. The precision of the computer word for recording numbers has important consequences for both arithmetical operations and for data storage.

Many people do not appreciate that use of computer variables and arrays having insufficient precision can lead to serious errors in calculations, particularly when results are required that must be obtained by subtracting or multiplying two large numbers. For example, the 'shorthand' method of estimating the variance of a set of numbers involves adding all the numbers together, squaring the result and dividing by the number of numbers. This 'constant' is then subtracted from the sum of the squares of all the numbers to obtain the sum of squared deviations. Box 9.2

explains that when many large numbers are involved there will almost certainly be large rounding errors occurring when the number of bits in the computer word is insufficient to handle the precision required.

Rounding errors Rounding errors are unlikely to be a problem when performing statistical calculations in large computers when the programming language allows double precision variables and arrays to be defined. They used to be troublesome in 16-bit micro computers, particularly if 'shorthand' methods of calculation were used. In the above example, it is much wiser to first calculate the average of the set of numbers, then to calculate the deviation of each number from the average and then sum the squared deviations. This method of estimating the sums of squares is not only closer to the original method of defining variation, but avoids rounding errors in the subtraction process.

In many systems used for image analysis data are coded as integers. The problems of accurately representing the areas and perimeters of polygons in raster format were noted in Chapter 3. Franklin (1984) explored the problem of data precision for other GIS operations, such as scaling and rotation, when the results of arithmetical operations are truncated to the nearest integer. As Figure 9.1*a* shows, scaling a simple triangle by a factor of three results in the point P being moved outside the triangle. Rotating point P (Figure 9.1*b*) moves it inside the circle.

The obvious way to avoid this problem is to increase the precision with which the computer represents numbers, i.e. to work with real numbers with a decimal representation.

As Franklin demonstrated, this merely pushes the problem to another level; it does not go away. The problem is one that is intimately linked with the way the computer represents numbers. It is possible to find real numbers for which computer implementations of simple arithmetic violate important real number axioms of distributivity, associativity, and commutativity.

For example, associativity:

$$(A + B) + C = A + (B + C)$$

This rule is violated in a computer that stores fewer than 10 significant digits for $A = 1.E10$, $B = -1.E10$, $C = 1$. Franklin showed that these problems can be corrected by using different methods of computation, which themselves bring extra problems of complexity and the need to develop or use special subroutines for arithmetical operations.

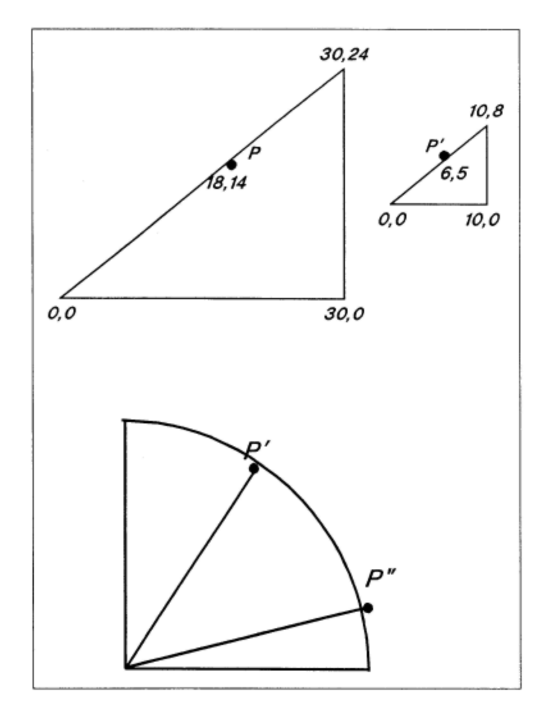


Figure 9.1. With integer arithmetic, scaling or rotation can cause points near boundaries to be rounded off inside or outside a polygon

Geographical coordinates and precision Chrisman (1984*b*) examined the role of hardware limitations on another problem in geographical information systems, namely that of storing geographical coordinates to the desired level of precision. Whereas 16-bit machines have presented few problems for storing the coordinates of low-resolution, single scene LANDSAT images, the high accuracy required by cadastral systems, or the sheer range of coordinates required to cover a continent result in numbers that are too large to be recorded in a single 16-bit computer word and 32-bit words or even 64 bits are necessary (Table 9.1). Fortunately this is no longer a serious problem. The 32-bit word used in many computers currently used for GIS, allows spatial dimensions to be recorded with the following precision:

Maximum dimension (metres)	Maximum precision attainable
10 000.00	dddd.dx
100 000.0	ddddd.x
1 000 000	dddddx

where d means good data, and *x* is the excess precision needed to avoid most of the topological dilem-

Table 9.1. The relation between computer word length and digital range and precision

Variable type	Number of significant digits (decimal)	Approximate decimal range
16-bit integer (2 bytes)	4	$-32768 \leq x \leq +32767$
Short real 32 bits (single precision 4 bytes)	6–7	$-3.37 \times 10^{38} \leq x \leq -8.43 \times 10^{-37}$ through true zero to $8.43 \times 10^{-37} \leq x \leq +3.37 \times 10^{38}$
Long real 64 bits (double precision 8 bytes)	15–16	$-1.67 \times 10^{308} \leq x \leq -4.19 \times 10^{-307}$ through true zero to $4.19 \times 10^{-307} \leq x \leq 1.67 \times 10^{308}$

mas of the kind shown in Figure 9.1. While it is unlikely that a user will require a precision better than 10 m for an area of 1000 × 1000 km, data from sensors like the French satellite SPOT with its 10 m resolution, which may be used to supply data for the resource inventory, mean that the 32-bit floating point representation in the GIS is stretched to the limit. Moreover, it may be necessary in the inventory to refer to ground control points that have been located with much greater precision.

Chrisman (1984*b*) and Tomlinson and Boyle (1981) have pointed out that locational precision is critical when the user wishes to interface different kinds of data sets that have been acquired at different scales and to different levels of precision. These problems are greater when working with established inventories that may have been geometrically using old 16-bit systems than when all data must be collected for specific projects, because often in the latter case, the data are collected from scratch.

Through national and international agreements and improvements in hardware and software, information on the quality of digital data is becoming an important part of the data itself.

Faults stemming from assumptions concerning the exactness of spatial entities

As already noted, most procedures commonly used in geographic information processing assume implicitly that (*a*) the source data are uniform, (*b*) digitizing procedures are infallible, (*c*) map overlay is merely a question of intersecting boundaries and reconnecting a line network, (*d*) boundaries can be sharply defined and drawn, (*e*) all algorithms can be assumed to operate in a fully deterministic way, and (*f*) class intervals defined for one or other 'natural' reason are necessarily the best for all mapped attributes. These ideas result from the traditional ways in which data were classified and mapped. They have presented large technical difficulties for the designers of geographical information systems but rarely have these problems been looked at as a consequence of the way in which the various aspects

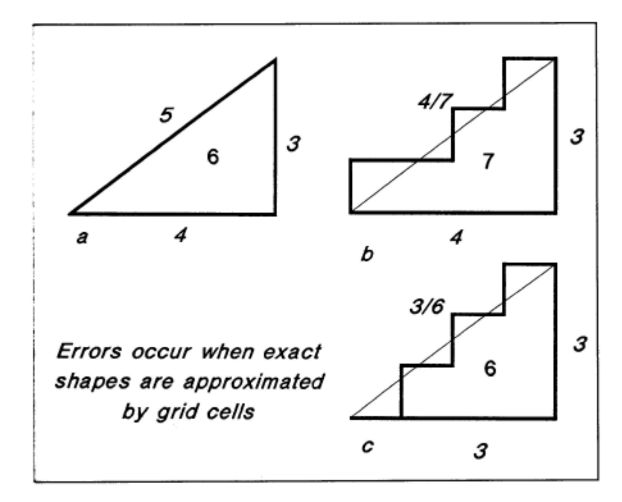


Figure 9.2. Errors occur when exact shapes are approximated by grid cells

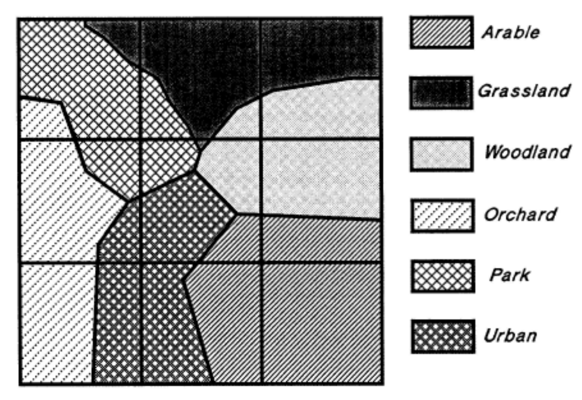


Figure 9.3. The mixed pixel problem occurs when grid cells are too large to resolve spatial details

of the world have been perceived, recorded, and mapped.

Many operations in geographical information processing require one or more spatial networks to be combined. The spatial networks may be composed of lines, of regular grids, or of irregular polygons. The overlay may be for the purposes of data conversion, such as converting a vector representation of a polygon net to raster form by overlaying a grid of given resolution, or for the purposes of data combination or modelling, such as when two polygon networks are intersected, or when the boundary of a watershed is used to cut out areas from an overlay of administrative units, or when data from soil polygons are input to a crop yield model.

This section covers the errors that can result from (*a*) converting spatial entities such as polygons from a vector to a raster representation, and (*b*) from overlaying and intersecting two polygon networks under the assumptions that spatial entities are exactly defined.

Errors resulting from rasterizing a vector map

Grid cells are only approximations? As Figure 9.2 shows, converting a vector triangle to unit pixels results in a serious loss of information. The area of the triangle should be 7 units, but could be taken to be 6 or 7 units depending on how the cells and their sides are counted. The hypotenuse could be 7 cell sides long if 4 cells are taken as an approximation of the diagonal, but only 6 if we opt for 3 cells—both are overestimates. Today, approximation errors with rasters are less of a problem because we have much larger and faster computer storage. In cartographic applications such as digital orthophoto maps (Plate 1) and on many laser and ink jet printers the grid cell is much smaller than the finest line drawn by a pen on a vector plotter—in fact, most plotters used today use raster technology. Only when large grid cells are used as basic database entities need we consider the different accuracies of a vector and a raster representation of space.

Mixed pixels Errors can arise in two ways when spatial phenomena are represented by an array of grid cells. The first and most obvious source of error is the problem of 'mixed pixels'; because each grid cell can only contain a single value of an attribute it is only the mean value that is carried in the cell. In the original LANDSAT imagery, in which each cell had a size of some 80 × 80 m, the signature of the pixel was a mean value of the reflectance averaged over the area of the whole cell; for SPOT the cells are 20 × 20 m so the spatial averaging is less. The differences in cell size mean that if part of the LANDSAT cell covered a highly reflecting surface such as water, this could so weight the mean reflectance as to give an over-representation of the area of 'water' compared to SPOT which might record other land cover types within the 80 × 80 m area. These kinds of classification error can occur whenever the size of the grid cell is larger than the features about which information is desired. It is a problem particularly when large-area grid cells are used to record many features in a complex landscape (Figure 9.3). In vector-raster conversion, the mixed

pixel problem leads to the dilemma of whether to classify a cell according to the class covering the geometric reference point of the cell (the centre or the south-west corner) or according to the dominant class occurring in the cell. In remotely sensed and other scanned imagery, the problem is complicated because the cell value is a weighted average of the information reaching the sensor from the area not only covered by the pixel but from nearby surrounding areas.

Vectors to fine rasters Converting polygons from vector to raster representation when using grid cells smaller than the polygons (see Chapter 4) brings with it the problem of topological mismatch when the smooth polygon boundaries are approximated by grid cells. Although high-quality raster scanners and plotters have largely removed the problem of loss of information through rasterizing from the visualization area of GIS, there are still many instances where thematic data, originally in vector polygon form, need to be rasterized to match data on regular grids, such as those obtained by remote sensing, or for some of the analysis examples in Chapter 7. Therefore it is necessary to estimate the seriousness of the problems of mismatch caused. Piwowar *et al.* (1990) examined several algorithms for vector to raster conversion for the quality of the results, the accuracy, the lateral displacement of boundaries, and their speed of operation. They concluded that not all algorithms worked equally well; some are fast but cause distortion, while others take more time but produce better results.

Note that in the following discussion of vector to raster conversion, the polygons are regarded as exact entities with precisely located boundaries; the errors of conversions are therefore merely the result of representing a geographical area by one geometry or another. Errors of misidentification or of the inability to define exactly what the area comprises are not treated here.

Statistical approaches to estimating the errors of vector to raster conversion Frolov and Maling (1969) considered the problem of error arising when a grid cell is bisected by a 'true' boundary line. They assumed that the boundary line could be regarded as a straight line drawn randomly across a cell. The mean square area of the cut-off portion of each bisected boundary cell i (the error variance) can be estimated by

$$V_i = aS^4 \qquad 9.1$$

where V is the error variance, S is the linear dimension of the (square) cell, and a is a constant. Frolov and Maling calculated the value of a as 0.0452 but subsequent work reported by Goodchild (1980) suggests that a better value is $a = 0.0619$.

The error variance in an estimate of area for any given polygon is given by a summation of all the errors from all the bounding cells. If m cells are intersected by the boundary, the error variance will be

$$V = maS^4 \qquad 9.2$$

with standard error

$$SE = (ma)^{1/2}S^2 \qquad 9.3$$

assuming that the contributions of each cell are independent. Goodchild (1980) suggests that this assumption should not always be regarded as valid.

The number of boundary cells m, can be estimated from the perimeter of the polygon. Frolov and Maling (1969) showed that m is proportional to $\sqrt{N}$, where N is the total number of cells in the polygon. The standard error of m is estimated by $kN^{1/4}a^{1/2}S^2$. Because the estimate of polygon area $A = N.S^2$, the standard error as a percentage of the estimate is proportional to $N^{-3/4}$ (Goodchild 1980), i.e.

$$SE = ka^{1/2}N^{-3/4} \qquad 9.5.$$

If the variable is cell side S instead of cell number N, the percentage error depends on $S^{3/2}$. Goodchild (1980) reports studies that have verified these relationships empirically.

The constant k depends on the polygon shape, long thin shapes having more boundary cells than a circular form of the same area. Frolov and Maling (1969) give values of k for various standard shapes, using the independent straight line hypothesis.

Switzer's method Switzer (1975) presented a general solution to the problem of estimating the precision of a raster image that had been made from a vector polygon map. His analysis does not deal with observational or location errors, but assumes that error is solely a result of using a series of points located at the centres of grid cells to estimate an approximate grid version of the original map. Switzer's method deals essentially with ideal choropleth maps, i.e. thematic maps on which homogeneous map units are separated by infinitely thin, sharp boundaries. The method assumes that a 'true' map exists, against which an estimated map obtained by sampling can be compared. Realizing that the 'true' map is often unknown or unknowable, Switzer showed that by applying certain assumptions and by using certain summary

BOX 9.3. SWITZER'S METHOD

The $P_{ij}(n^{-1/2})$ and $P_{ij}(2n^{-1/2})$ probabilities are estimated from a frequency count as follows:

1. Estimate the total number of cell pairs at distance $d = 1$ cell width. The total number of pairs at a given distance is equal to

$$NPAIRS = 4^*(P^*Q) - 2^*d^*(P + Q)$$

where P = number of rows, Q = number of columns in the grid, and the second term is a correction for the cells on the edge of the grid.

2. For each pair of mapping units i and j, count the number of cell pairs along and up and down the grid that lie in different mapping units ($TALLY_{ij}$).

3. Compute $P_{ij}(n^{-1/2})$ as $TALLY_{ij}/NPAIRS$.

4. Repeat steps 1–3c with $d = 2$ cell widths to estimate $P_{ij}(2n^{-1/2})$.

5. Calculate O_{ij} from equation (9.5).

6. Calculate total mismatch for each mapping unit O_i as the sum of the O_{ij}s, remembering that the mismatch of $O_{ji} = O_{ij}$.

7. Calculate total mismatch as the sum of the O_is.

statistics, errors of mismatch could be estimated from the estimated or gridded map itself.

The analysis begins by assuming that a map M has been partitioned into k homogeneous map units, or colours. Each of the k map units may be represented on the map by one or more sub-areas. This 'true' map is estimated by laying an array of n basic sampling cells over the map. Here we shall only consider the situation where the array of sampling cells is regular and congruent, and each cell is defined by a single sampling point at the cell midpoint. The map units on the 'true' map are denoted M1, M2, . . . Mk, and on the estimated map by M1, M2, . . . Mk. Each cell on the estimated map is allocated to a map unit Mi if the sampling point in the cell falls within map unit Mi on the 'true' map. This is the procedure commonly used when converting a vector polygon network to raster format. For the purposes of this analysis we shall, like Switzer, assume the total area of the map is scaled to unity, i.e. $A(M) = 1$.

The degree of mismatch of the estimated map is a function of two independent factors, (*a*) the complexity of the true map, and (*b*) the geometrical properties of the sampling net. Considering first the complexity of the map, we can define a quantity $P_{ij}(d)$ as the probability that a random point is in true map unit i and that the cell centre point is in true map unit j when the points are separated by distance d. Switzer derived the following expression for the percentage overlap O_{ij} for each pair of rasterized map units i,j using square grid cells,

$$O_{ij} = 0.60P_{ij}(n^{-1/2}) - 0.11P_{ij}(2n^{-1/2}) \qquad 9.5$$

(Note that the values of the coefficients differ from Switzer's published formula; the corrections are given by Goodchild (1980)).

The total error for all k map units is given by

$$O = \sum_{i=j}^{k} O_{ij} \qquad 9.6$$

Box 9.3 shows how O can be estimated in practice.

An example. Figure 9.4*a* shows the boundaries of a simple thematic map depicting soil or geological units. Assuming that they form a true map, what will be the relative mismatch errors arising from digitizing it using grid rasters of different sizes, as shown in Figures 9.4*a*,*b* for 16 × 16 and 32 × 32 grids, respectively?

Table 9.2 gives the results. For a grid measuring 16 × 16 cells, there are 960 cell pairs at distance $d = 1$. The total number of cell pairs straddling a boundary lead to frequency estimates for each mapping unit. For a distance $d = 2$, the number of pairs is 896. Entering the frequency estimates into equation (9.6) leads to an estimated mismatch of 9.5 per cent. Using the 32 × 32 cell grid leads to an estimate of 4.1 per cent,

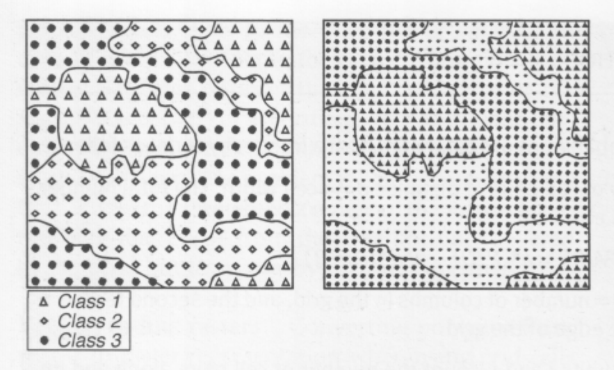


Figure 9.4. Rasterizing a vector map at two grid sizes to estimate rasterizing errors

demonstrating that a factor 4 increase in the number of grid cells is needed to reduce the estimation error by half. Both these estimates of mismatch compare favourably with the estimates of mismatch obtained by measuring the areas of the mapping units on the original map. Note that this means that mismatches with a printer of 600 dpi are approximately half those of one using 300 dpi.

Bregt *et al.*'s method Bregt *et al.* (1991) developed an elegant method for estimating the error associated with vector-raster conversion called the *double-conversion method*, because it involves rasterizing the map twice. First the vector to raster conversion is carried out using the desired target raster size; this produces what they call the base raster. The map is then rasterized to a very much smaller grid and the two are compared. Those cells in the fine raster differing from those on the base raster provide an estimate of the error in the base raster.

Bregt *et al.* compared the errors so obtained with a parameter called the *boundary index* (*BI*), which is defined as the boundary length in centimetres per square centimetre of the map. The *BI* is calculated by dividing the total length of the polygon boundaries by their total area. They found that for a given cell size the rasterizing error (as a percentage mismatch) is a linear function of *BI*. They distinguish two situations, (*a*) that where the cell on the base raster is classified according to the polygon in which its central point falls, and (*b*) cell classification by the polygon that dominates its area. Table 9.3 presents the results.

Bregt *et al.* compared their method with Switzer's and demonstrated that it provides easier and better estimates of the rasterizing error since only the *BI* needs to be computed. *BI* values are independent of the units used. The disadvantage is that the regression equations need to be worked out for all possible situations, whereas Switzer's method is completely general and requires no previous work.

Errors associated with digitizing a map, or with geocoding

As already noted, the methods of Switzer, Goodchild, and Bregt *et al.* to estimate mismatch assume implicitly that a 'true' map exists that has homogeneous (uniform) mapping units, and infinitely sharp boundaries. In practice, however, even the best-drawn maps are not perfect, and extra errors are introduced by the digitizing process as authors such as Blakemore (1984), Bolstad *et al.* (1990), Dunn *et al.* (1990), and Poiker (1982) have pointed out. Consider the problem of boundary width and location (the problem of within-map unit homogeneity will be dealt with later) on a

Table 9.2. The results of using Switzer's method on the map in Figure 9.3

Estimates/grid size	16 × 16 grid			32 × 32 grid		
Mismatch per polygon pair	L_{12} 0.020	L_{21} 0.020	L_{31} 0.008	L_{12} 0.010	L_{21} 0.010	L_{31} 0.004
	L_{13} 0.008	L_{23} 0.014	L_{32} 0.014	L_{13} 0.004	L_{23} 0.006	L_{32} 0.006
Mismatch per polygon	L_1 0.028	L_2 0.034	L_3 0.022	L_1 0.014	L_2 0.016	L_3 0.011
Total mismatch (%)	8.46			4.11		

Table 9.3. Relations between rasterizing method, cell size, and rasterizing error

Rasterizing method	Raster cell size (mm)	Regression equation	Variance explained (%)
Central	1 × 1	$L = 2.4BI$	99.8
Dominant	1 × 1	$L = 2.4BI$	99.8
Central	2 × 2	$L = 4.7BI$	99.8
Dominant	2 × 2	$L = 4.6BI$	99.8
Central	4 × 4	$L = 9.0BI$	99.4
Dominant	4 × 4	$L = 8.7BI$	99.0

Source: Bregt *et al.* 1991

digital map of polygons in vector format. The digital map will almost certainly have been derived by digitizing a paper version of the map. There are two sources of potential error—(*a*) errors associated with the source map, and (*b*) errors associated with the digital representation.

(*a*) Apart from the potentially correctable errors of paper stretch and distortion in the printed map or source document, errors arise with boundary location simply because drawn boundaries are not infinitely thin. A 1 mm line on a 1 : 1250 map covers an area 1.25 metres wide; the same line on a 1 : 100 000 map covers an area 100 m wide. A detailed 1 : 25 000 soil or geological map measuring 400 × 600 mm may have as much as 24 000 mm of drawn lines covering an area of 24 000 sq. mm or 10 per cent of the map area! Common sense suggests that the true dividing line should be taken as the midpoint of the drawn line, but it is not being cynical to state that the area of the map covered by boundary lines is simply an area of uncertainty, and possibly, confusion. When these boundary lines are converted by digitizing, extra errors arise because with hand digitizing the operator will not always digitize exactly the middle of the line, and with scanners, errors will arise with the data reduction algorithms used.

(*b*) The representation of curved shapes depends on the number of vertices used (Aldred 1972: 5). Consequently, the relative error of digitizing straight lines is much less than that resulting from digitizing complex curves. Translating a continuous curved line on a map into a digital image involves a sampling process: only a very small proportion of the infinity of possible points along a curve is sampled (see in Figure 9.5; Smedley and Aldred 1980).

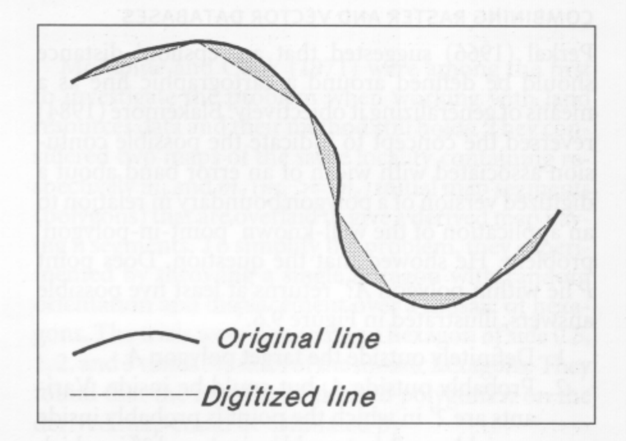


Figure 9.5. Digitizing a line is a sampling process

Clearly, boundaries on thematic maps should not be regarded as absolute, but as having an associated error band or confidence interval. MacDougal (1975) suggested that the total boundary inaccuracy could be estimated by

$$H = \sum_{i=1}^{N} (h_i l_i)/T \quad 9.7$$

where h_i is the horizontal error (in standard deviations) of line i, length l_i, N is the number of boundary lines, and T is the total area of the map. If all boundary lines are the same type (e.g. they are all soil boundaries or all land use boundaries) equation (9.7) simplifies to

$$H = (hL)/T \quad 9.8$$

The total line length L was originally estimated by placing a grid over the map and counting the number of crossings, K, and using the formula

$$L = (TK)/0.6366 \qquad 9.9$$

where 0.6366 is a constant described by Wentworth (1930), but today the total length can easily be computed from the database.

In an empirical study, Bolstad *et al.* (1990) report that the errors due to the manual digitizing of 1 : 25 000–1 : 50 000 soil maps were quite small, and for the United States, of less importance than positional errors due to uncertainty in georegistration.

Error bands around a digitized line: problems for point-in-polygon searches and when combining raster and vector databases

Perkal (1966) suggested that an 'epsilon' distance should be defined around a cartographic line as a means of generalizing it objectively. Blakemore (1984) reversed the concept to indicate the possible confusion associated with width of an error band about a digitized version of a polygon boundary in relation to an application of the well-known 'point-in-polygon' problem. He showed that the question 'Does point *P* lie within polygon *A*?' returns at least five possible answers, illustrated in Figure 9.6.

1. Definitely outside the target polygon *A*.
2. Probably outside *A*, but could be inside. Variants are 2′ in which the point is probably inside a neighbour *B*, but could be in *A*, and 2″ in which the point is probably outside *A* but could be in either of two neighbours *B* or *C*.
3. On the boundary—indefinite.
4. Probably inside *A*, but could be outside; other variants are 4′ where the point is probably in *A* but could be in a neighbour *C*, and 4″ in which the probably 'in' point could also be in one of two neighbours *B* or *C*.
5. Definitely in *A*.

'Definitely in' records the core area within the error band; 'possibly in' records a point that falls within the overlap of the inner half of the confidence band and the polygon. 'Possibly out' records a point that falls in the outer half of the confidence band; technically speaking the point would be returned as falling outside the polygon, but it could actually be inside the 'true' polygon if it had been erroneously digitized or geocoded. An ambiguous point has coordinates that coincide exactly with a point on the digitized boundary—such points are rare, but do occur.

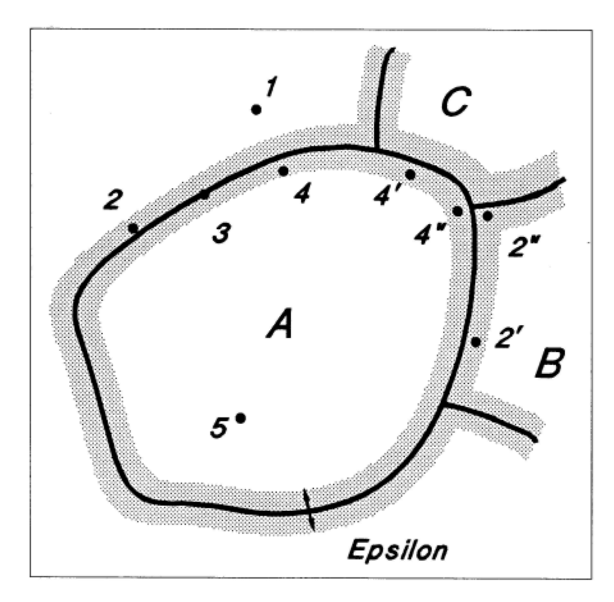


Figure 9.6. Perkal's concept of an epsilon error band around a digitized line

Blakemore (1984) illustrated the effects of these kinds of errors when dealing with problems of combining a vector polygonal net with a square grid cell network. The problem he chose was that of overlaying a UK Department of Industry 1 km square grid data base of industrial establishments on a polygonal map of 115 North-West England employment office areas. A total of 780 entries in the database geocoded to a 1 km square grid resolution were tested for their inclusion in the polygon network. The 1 km square grid leads to an epsilon or confidence band of 0.7071 km. Table 9.4 presents Blakemore's results.

The 'possibly out definite' class includes data points that fell outside the polygon network of employment office areas altogether. 'Possibly in' refers to points falling within the inner half of the error band in polygons on the edge of the polygon net. 'Unassignable' refers to points that fell outside the error band of the outer boundaries of the outer polygons. In some circumstances the point-in-polygon routine suggested that the industry was located in the sea! 'Possibly in/out 2 polys.' refers to points that were flagged as being possibly in and possibly out of two adjacent polygons; 'possibly in/out > 2 polys' refers to points that were possibly in or out of more than 2 polygons. 'Ambiguous' refers to those points actually occurring on the digitized polygon boundaries. The implication of the study was that only 60 per cent of the workforce in the industries in the database could definitely be

Table 9.4. Epsilon error results

Category	%
Possibly out definite	1.5
Possibly in	4.4
Unassignable	1.4
Possibly in/out 2 polys.	29.8
Possibly in/out > 2 polys.	6.7
Ambiguous	1.2
Subtotal	45.0
Definitely in	55.0
Total	100.0

associated with an employment office area. The mismatch errors and ambiguities were relatively larger for long, thin polygons and for employment areas having narrow protuberances or insets than for large, broadly circular areas. The study resulted in a considerable amount of validation and checking of the data bases to ensure that the errors brought about by the grid-cell point geocoding were removed. Perkal's epsilon assumes the boundary is real, the problem merely being one of knowing its location. Sometimes it is not the location but the *existence* of the boundary that is in doubt, and then other methods must be used—see Chapter 11.

Errors associated with overlaying two or more polygon networks

Spatial associations between two or more thematic maps of an area are commonly displayed or investigated by laying the polygonal outlines on top of one another and looking for boundary coincidences. Before the days of digital maps, the process was achieved using transparent sheets, and the boundary coincidences were established using fat marker pens to trace the results. The onset of the digital map promised better results because all boundaries were supposed to be precisely encoded, but in fact the result of the new technology was to throw up one of the most difficult and most researched problems in computer cartography. Not only did a solution of the problem in technical terms cost many years' work but investigations have shown that the results of overlay throw up more questions about data quality and boundary mismatching than they solve.

McAlpine and Cook (1971) were among the first to investigate the problem when working with land resources data and their method still holds. They considered two maps of the same locality containing respectively m_1 and m_2 ($m_1 >= m_2$) initial map segments (polygons) that are overlaid to give a derived map having n segments. To simplify the problem, they experimented by throwing a single hexagon with random orientation and displacement over a mosaic of hexagons. The trials were done using a hexagon of side 0.5, 1, 2, and 3 times the sides of the mosaic hexagons. They found that the number of derived polygons n on the derived map could be estimated by

$$n = m_1 + m_2 + 2 \cdot \{m_1 m_2\}^{1/2} \qquad 9.10$$

for two maps, which for k maps can be generalized to

$$n = \left[\sum_{i=1}^{k} m_i\right]^2 \qquad 9.11$$

McAlpine and Cook (1971) showed that map overlay gave rise to a surprisingly large proportion of small polygons on the derived map. They applied their analysis to a case-study of overlaying three maps of scale 1 : 250 000 of census divisions, present land use intensity and land systems from Papua and New Guinea containing 7, 42, and 101 initial polygons respectively. The overlay of the three maps gave 304 derived polygons (Equation (9.11) estimates 368 derived polygons, but McAlpine and Cook regard this as satisfactory). The overlay process resulted in 38 per cent of the area being covered by polygons having areas of less than 3.8 sq. kilometres.

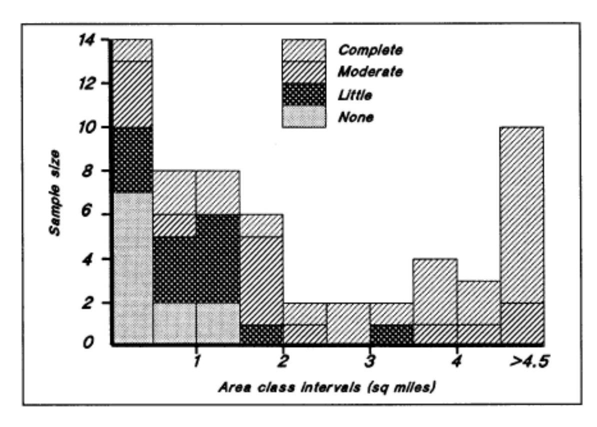


Figure 9.7. Measure of agreement between initial and derived polygon descriptions after polygon overlay

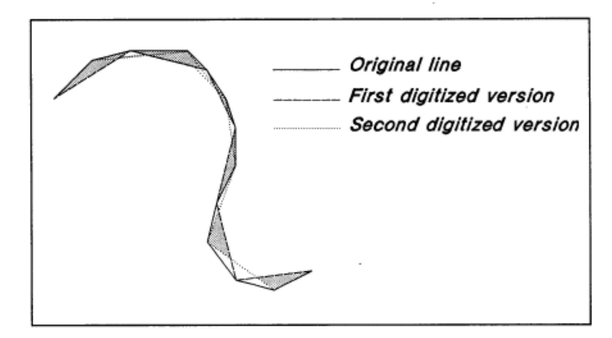


Figure 9.8. How spurious polygons occur in map overlay when the same line is digitized twice

The results of the overlay were evaluated by classifying the derived polygons by size and boundary complexity (i.e. polygons bounded solely by initial mapping segments, those bounded only by land use and land system boundaries, and those bounded by all three types of boundaries). A 10 per cent random sample of derived polygons was evaluated by three colleagues to determine the measure of agreement between the initial and the derived polygon descriptions. As Figure 9.7 shows, the lack of agreement was substantial for the smallest derived polygons, and some 30 per cent of the area of the derived map was represented by polygons that had little or no agreement with the initial descriptions.

Goodchild (1978) extended the discussion of the polygon overlay problem to show that the number of derived polygons is more a function of boundary complexity than the numbers of polygons on the overlaid maps. He showed that an overlay of two polygons having respectively v1 and v2 vertices could produce any number of derived polygons from three to v1.v2 + 2 when all Boolean operations including .NOT.A.AND.NOT.B are used. Moderate numbers of derived polygons are produced when, as in McAlpine and Cook's example, the overlaid maps show statistical independence. When the boundaries of polygons on the source maps are highly correlated, however, serious problems arise through production of large numbers of small, 'spurious' polygons. Prominent and important features, such as district boundaries or rivers, may occur as part of polygon boundaries in several maps. These several representations of the same boundary will have been separately digitized, but because of digitizing and other errors will not exactly coincide.

The spurious polygon problem contains two apparent paradoxes. First, the more accurately each boundary is digitized on the separate maps, and the more coordinates are used, the larger the number of spurious polygons produced. Second, subjective methods of map drawing, designed to assist the eye in generalizing when using manual methods of overlay, result in large problems when working with digital maps.

Goodchild (1978) analysed the situations in which spurious polygons were most likely to occur through the conjunction of two digitized versions of the same arc, with n_1 and n_2 vertices respectively (Figure 9.8). Goodchild, using the statistics of runs of binary symbols, showed that the number of spurious polygons S generated by conjunction of two arcs having n_1 and n_2 vertices ranges from

$$Smin = 0 \qquad 9.12$$

to

$$Smax = 2\min(n_1,n_2) - 4 \qquad 9.13$$

with a random expectation of

$$E(S) = [2n_1,n_2/(n_1 + n_2)] - 3 \qquad 9.14$$

if symbols occur randomly in sequence along the conjoined arcs. The minimum value of S occurs when the overlap is of maps having symbols of one type occurring together; the maximum value of S occurs for maximum intermixing. By simulating five possible situations in which arcs were conjoined, Goodchild showed that equation (9.14) overestimates the average number of spurious polygons that can occur by some 17 per cent. The actual number of spurious polygons found never exceeded 71 per cent of Smax. The more carefully a map is digitized, however, the larger the values of n_1 and n_2, and so the larger the number of spurious polygons will become.

Spurious polygons are in fact equivalent to the mismatch areas resulting from rasterizing a polygon. Their total area should decrease as digitizing accuracy increases, but the greater problem is their removal to avoid nonsense on the final map. They can be removed by erasing one side on a random basis, after screening the polygon for minimum area, or the two end-points can be connected by a straight line and both sides dissolved. A more sophisticated approach is to consider all points within a given distance from the complex arc as estimates of the location of a new line, and then fit a new line by least squares or maximum likelihood methods. Unless one version of the digitized boundary can be taken to be definitive, it is highly likely that the complex line will be moved from its topographically 'true' position. The net result of overlaying a soil map (having not very exact boundary locations) with a county boundary map (topographically exact boundaries) may be that the topographic boundaries become distorted unless the user specifies that they should remain constant.

Adopting exact paradigms of exact boundaries or smooth contour lines for spatial entities presents problems for converting from one representation to another and for entity overlay and intersection, but methods exist to estimate the errors that are involved in these actions. The degree of error caused by forcing of spatial phenomena into possibly inappropriate, exact, crisply defined entities has received less attention but may be a major source of unseen errors and information loss. Geographical phenomena with uncertain boundaries are covered in Chapter 11. (See also Burrough and Frank 1996.)

Summary: errors and mistakes

As in any manufacturing process, poor quality raw materials leads to poor quality products. Spatial information systems, however, also make it possible to turn good raw materials into poor products, if proper attention is not paid to the ways data are collected, modelled, and analysed. Conventionally, data quality has been linked to the precision of geographic coordinates, but today, exactness of location is but one aspect of data quality. The reader should also be aware that sometimes people expect a higher-quality product than is strictly possible, or even necessary. For example, for auto navigation it is extremely important that the database is geometrically and factually precise, but for marketing studies (e.g. Plates 2.5–2.8) extreme spatial accuracy is not only necessary, but threatens individual privacy.

Questions

1. Review the different methods that can be used to determine errors in spatial data. Consider a range of different GIS applications and assign appropriate error analysis techniques to each application.

2. Design a meta data system for recording the results of data quality and error propagation as active aspects of a spatial data set.

3. Review four practical situations where lack of information about errors could be critical for the acceptance of the results of GIS analyses.

4. Compile lists of the sources of errors for each of the examples of GIS analysis given in Chapters 6 and 7 and classify these errors using the terms given in this chapter. For

each example decide which source of error is most likely to be critical for successful analysis.

Suggestions for further reading

David, B., van den Herrewegen, M., and Salgé, F. (1996). Conceptual models for geometry and quality of geographic information. In P. A. Burrough and A. U. Frank (eds.), *Geographic Objects with Indeterminate Boundaries.* Taylor & Francis, London.

Goodchild, M., and Gopal, S. (1989). *The Accuracy of Spatial Databases.* Taylor & Francis, London.

Guptill, S., and Morrison, J. (eds.) (1995). *The Elements of Spatial Data Quality.* Elsevier, Amsterdam.

Thapa, K., and Bossler, J. (1992). Accuracy of spatial data used in Geographic Information Systems. *Photogrammetric Engineering and Remote Sensing,* 58: 841–58.

TEN

Error Propagation in Numerical Modelling

This chapter reviews the sources of statistical uncertainty in spatial data, particularly with reference to the creation and use of continuous fields and the propagation of errors through numerical models. Following an exposition of how errors can accrue from spatial mismatch between support size, sampling interval and spatial correlation structures, the chapter presents both theory and tools for computing error propagation. Monte Carlo methods for error analysis are followed by the statistical theory of error propagation. Comparative studies of interpolation errors and errors propagated through regression models linked to simple cost-benefit analysis show users how to choose the best techniques for predicting heavy metal pollution in different circumstances of data abundance. Given knowledge on errors and error propagation, an intelligent GIS would be able to check the intrinsic quality of the results of GIS modelling and advise users how best to achieve the precision of the result they require.

Statistical approaches to error propagation in numerical modelling

In much quantitative environmental modelling with GIS, the attributes attached to points, lines, polygons, or grid cells are the inputs to numerical models for computing values of derived attributes which become a new attribute of the polygon or cell. Very often, the calculation is assumed to produce an exact result because most GIS provide no means to examine the effects of errors in the input data on the result of the model. However, it is clear that the quality of the results of quantitative models depends on three factors, namely

(*a*) the quality of the data,
(*b*) the quality of the model, and
(*c*) the way data and model interact.

To get reliable results it is very important to know how uncertainties in both model parameters and data propagate through the models. To analyse error

propagation (i.e. how uncertainties accumulate and affect the end-result) we need

(*a*) sources of error estimates,
(*b*) error propagation theory, and
(*c*) error propagation tools.

Sources of statistical uncertainty

Sources of statistical uncertainty in data include measurement errors and spatially correlated variation that cannot be explained by physical models. Chapter 6 showed how methods of geostatistical interpolation and conditional simulation (*stochastic imaging*) can be used to generate error surfaces for interpolated data. The simplest way to store data on errors in a GIS is to assume that all data are normally distributed and that the error is correctly expressed by the standard deviation. This means that all attributes for an entity or a grid cell should be expressed by two numbers, the recorded or mean value and its standard deviation; this of course, automatically leads to a doubling of the space required to store the attribute data. Linking GIS to statistical packages (Chapter 7) may also be used to compute statistical parameters of the data.

Effects of mismatches in spatial correlation structures An important, but unseen source of uncertainty in spatial data, may be the size of the support used to collect the data, particularly when data from different sources must be combined in the GIS. Even if the data share the same grid, different sets of spatial data can be poorly matched, however, when their spatial patterns have been sampled at a different resolution. Simply overlaying polygons or bringing vector and raster overlays or feature planes to the same geometric resolution does not guarantee that you can combine the spatial data from the different layers in a meaningful way. The belief that point observations are necessarily representative for larger areas when spatial variation is unseen is widespread in spite of many publications proclaiming otherwise (e.g. Burrough 1993, Beckett and Webster 1971, Openshaw 1977, Openshaw and Taylor 1979).

Mismatch in spatial and temporal correlation structures may be one of the greatest problems when combining data from different sources or from different phenomena or attributes, and may lead to much disillusionment with GIS. It is not GIS that is the problem but the way the data have been collected and aggregated.

Correlation structures are important!

If sample surveys for different attributes are tuned to different spatial correlation structures then it is difficult to obtain a sensible match for spatial modelling. The problem is not with GIS, but with the balance of spatial variation in natural phenomena. Analysis of indices of spatial covariation such as the variogram can be useful to ensure that different data are spatially compatible.

Mismatch can occur through:

(*a*) each phenomenon being measured on a different support (area/volume);
(*b*) each phenomenon has a different intrinsic spatial variability;
(*c*) some phenomena being sampled directly while others are collected or classified using externally imposed spatial aggregation blocks that are inappropriate;
(*d*) the spatial variation of different phenomena is governed by processes that operate at different scales. To complicate matters, the same attribute can have different spatial correlation structures at different scales.

The following example (Ten Berge *et al.* 1983) demonstrates how difficult it can be to relate contemporaneous measurements of *the same variable* to each other when measurements are made on different supports and these tune in to different spatial correlation structures. The aim of the study was to map the spatial variation of soil surface temperature across a 350 m wide, ploughed experimental field at the Ir. A.P. Minderhoudhoeve Experimental Farm in the Flevopolders, the Netherlands. The area is extremely flat, but because of differences in the patterns of sedimentation over the farm, annual thermal infra red surveys using airborne scanners carried out in early spring in the years 1980–2 detected effective surface temperature differences across ploughed experimental fields of some 0.6 degrees Celsius.

The aim of the fieldwork was twofold. First the investigators hoped to determine which property of the bare soil surface was responsible for the temperature variations over the field, ultimately to determine if

the surface temperature variations were causally linked to crop response. The second aim was to relate the airborne measurements of reflected radiation to ground-based measurements which are not only unaffected by atmospheric absorption but are cheaper and easier to carry out at any time.

The experiment was carried out under clear weather conditions at midday on 26 March 1986. Thermal infra-red imagery was obtained by flying a Daedalus DS 1260 multispectral scanner using the 8–14 μm spectral window (Daedalus Enterprises Inc., PO Box 1869, Ann Arbor, Michigan 48106, USA) in a light aircraft (Plate 4.2). The altitude of the aircraft carrying the scanner yielded a pixel size of 1.5×1.5 m, which means that all variations within a block of this size were averaged out to yield a single data value (a weighted average) for this support.

Ground data on reflected radiation and soil properties were collected at the same time as the aerial survey. For this comparison we use only the thermal infrared emission of the soil surface and information about the texture of the 0–5 mm surface layer of soil. The field data were collected at points spaced 4 m apart along two 200 m long transects (50 observation points) located perpendicular to the expected main temperature gradients. The transects were clearly marked so that they could easily be located on the airborne thermal imagery. The thermal infrared emission data for the soil surface were collected by a portable Heimann KT-15 radiometer mounted on a tripod with a spatial resolution of 0.03 m^2 (*c.* 0.17×0.17 cm) for the same 8–14 μm spectral window. Soil texture samples were obtained by bulking 9 subsamples taken within an iron grid measuring 1.00×1.00 m laid over the soil at each sample site, thereby effectively creating a support for the texture data of 1 square metre. In addition, surface soil temperatures for the 0–5 cm layer were measured with a needle thermocouple with digital readout.

Recapitulating, the survey provided data for two 50-point transects but the data had been collected using different support sizes. The airborne Infrared scanner data had a resolution of 1.5×1.5 m, the Heimann infrared data had a resolution of 0.17×0.17 m, and the soil texture data had an effective support size of 1.00×1.00 m. All data were first examined for normality, and none were found to be seriously non-normally distributed.

The effects of the different support sizes became evident when the different data were plotted to examine the spatial variation. Figure 10.1 shows the plots of the original data and their variograms which have been normalized to make comparison easier. Clearly the two infrared measurements show little relation to each other. The airborne data reflect the coherent spatial variation over the field but the Heimann data show only short-range variation or noise; a variogram model cannot be fitted to these data. Observations with the thermocouple demonstrated why this is so. The experimental field had been ploughed and had plough ridges of some 15–20 cm high some 25–30 cm apart. One side of the ridges was warmed by the low March sun to yield surface soil temperatures of 20–21 degrees Celsius; the other side of the plough ridges were in shadow and temperatures were only 14–15 degrees Celsius. Consequently the short range differences in temperature were causing the large fluctuations in the Heimann readings and swamped any weak variations across the field. Because the Daedalus scanner pixels were large enough to average out these short-range variations they returned information about the longer-range variations over the field. The soil texture data, having been bulked to a support similar in size to the airborne data shows a similar pattern and variogram to the Daedalus scanner data and suggests that the gradual variations in surface temperature across the field are linked to spatial variations in surface texture of the soil.

Lessons learned. The lesson provided by this example is that it is essential to know the size of the support that has been used to collect 'point' data or bulked samples. This information should be supplied as a matter of course in the metadata files describing the provenance and quality of publicly and commercially available databases (see Chapter 12), but most often this information is not easy to come by.

Error propagation theory (1): Monte Carlo simulation

Given that we are (*a*) aware of the possibility of statistical errors in the data, and (*b*) we have the means to quantify these errors using ordinary statistics or geostatistics, stochastic simulation or retrospective validation with independent data, the question arises as to how we can use this information to quantify and then reduce the inaccuracies that may accrue in the results of numerical modelling. Put simply, if a new attribute U is defined as a function of inputs A_1, $A_2, \ldots A_n$, we want to know what is the error associated with U, and what are the contributions from each A_n to that error? If we can solve this problem we can

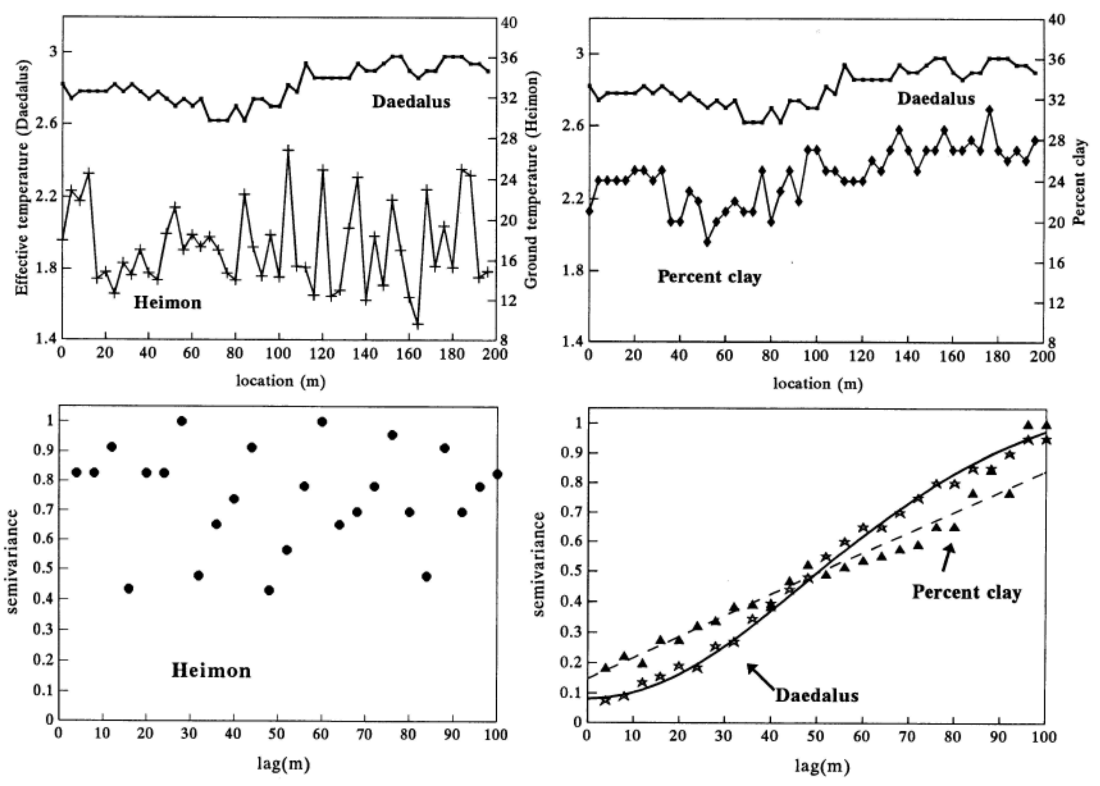


Figure 10.1. Top left: plots of effective soil surface temperature against location, as measured by the Daedalus airborne scanner and the Heimon ground instrument.
Top right: plots of Daedalus data and per cent clay for bulked samples from soil surface
Bottom left: experimental variogram of Heimon data
Bottom right: variograms of Daedalus data and per cent clay

attach a pedigree to the results of modelling and can compare the results of scenarios with confidence.

The simplest, but most computer intensive approach, to error propagation is to treat each attribute as having a Gaussian (normal) *probability distribution function* (PDF) with known mean μ and variance σ^2 for each entity or cell. In the simplest case we would use a single PDF for all cells, and we assume stationarity. If more information about spatial contiguity is available we can use conditional simulation to estimate cell-specific PDFs that reflect the location of known data points and the spatial correlation structure of the attributes (Chapter 6).

The arithmetical operation to derive new data is then carried out not with the single mean values but using a value that is drawn from the PDFs for each cell. To take care of the variation within the PDFs the calculations are repeated many times (at least 100 times) in order to compute the mean result per entity or pixel and its standard deviation. The technique is popularly known as the *Monte Carlo* method, because of its reliance on chance.

Although the Monte Carlo method is computer intensive (known as a 'brute force' technique) it provides interesting information about how possible errors in the data can affect the results of numerical operations in different parts of a geographic area. Consider the operations described in Chapter 8 for computing the derivatives of a raster DEM, such as slope. Elevation data are frequently accompanied by specifications of the RMS (root mean squared error) which is assumed to apply uniformly over the whole domain. The effect of different levels of RMS error in the DEM on calculations of the slope and the topological drainage net can be investigated as follows. An error field with mean $\mu = 0$ and variance σ^2 is generated for a given stand-

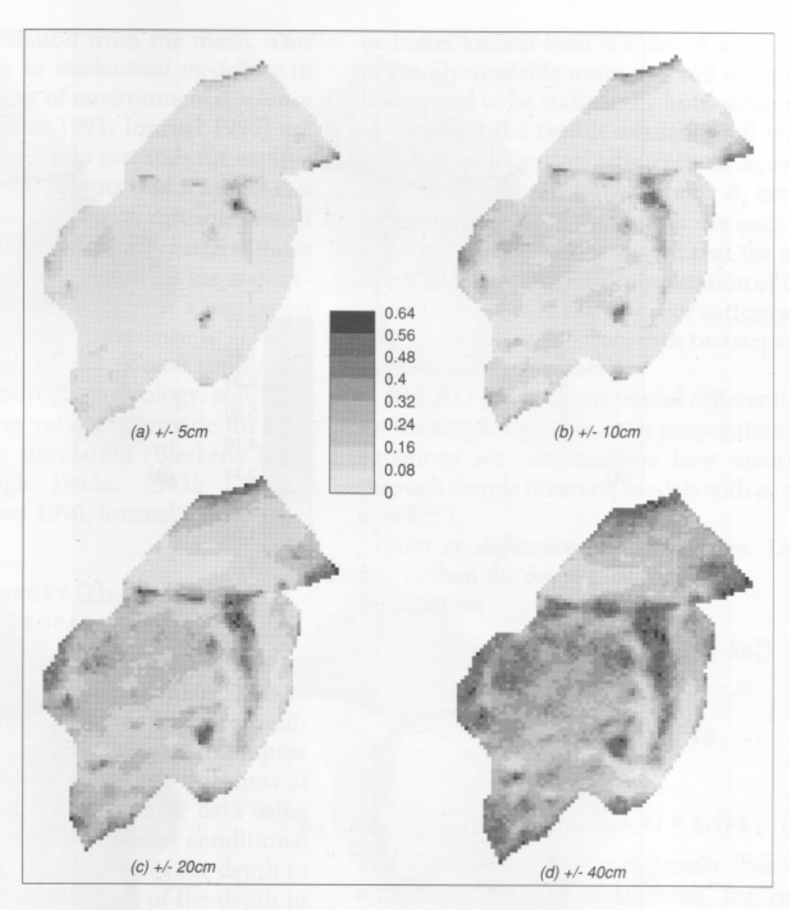


Figure 10.2. The effect of adding an RMS error to the DEM on the estimates of slope (degrees): (*a*) ± 5 cm, (*b*) ± 10 cm, (*c*) ± 20 cm, (*d*) ± 40 cm. Grey scale shows relative errors

ard deviation, say ± 1.0 m. This is added to the original DEM and the slope operation is carried out using the algorithm of choice. The result is stored and the calculation repeated 100 times, each time yielding a different realization of the derived slopes. Averaging the 100 results yields a mean slope map and a map of the standard deviation of the slope for each cell. Dividing the standard deviation map by the mean slope map yields a map of relative error. The procedure can be repeated for other values of σ^2.

Figure 10.2 presents results for the Catsop catchment (see Appendix 3) for RMS errors of 5, 10, 20, and 40 cm respectively. It clearly shows the strong effect of the relative errors in areas of low relief, implying that there it will be more difficult to obtain clear-cut results.

Not only slope, but other derivatives such as local drainage networks and upstream contributing catchments are also very sensitive to errors in the elevation data. Figure 10.3 presents the results of an analysis to examine how RMS errors in the DEM affect the size of the upstream contributing area. The procedure is simply to compute an error surface with a given RMS error and add this to the DEM, and then to compute the local drainage network and the upstream contributing area by the usual procedures such as the D8 algorithm (Chapter 8). Repeating the simulation 100 times and averaging for different levels of RMS error clearly shows how sensitive derived attributes such as contributing upstream area are to errors in elevation. In some parts of the area the drainage lines are strongly concentrated while in other areas the average cumulative upstream area suggests drainage is diffuse. Note that in this case even a small RMS error of ±5 or 10 cm (which is negligible in comparison to the surface roughness created by ploughing) can strongly affect

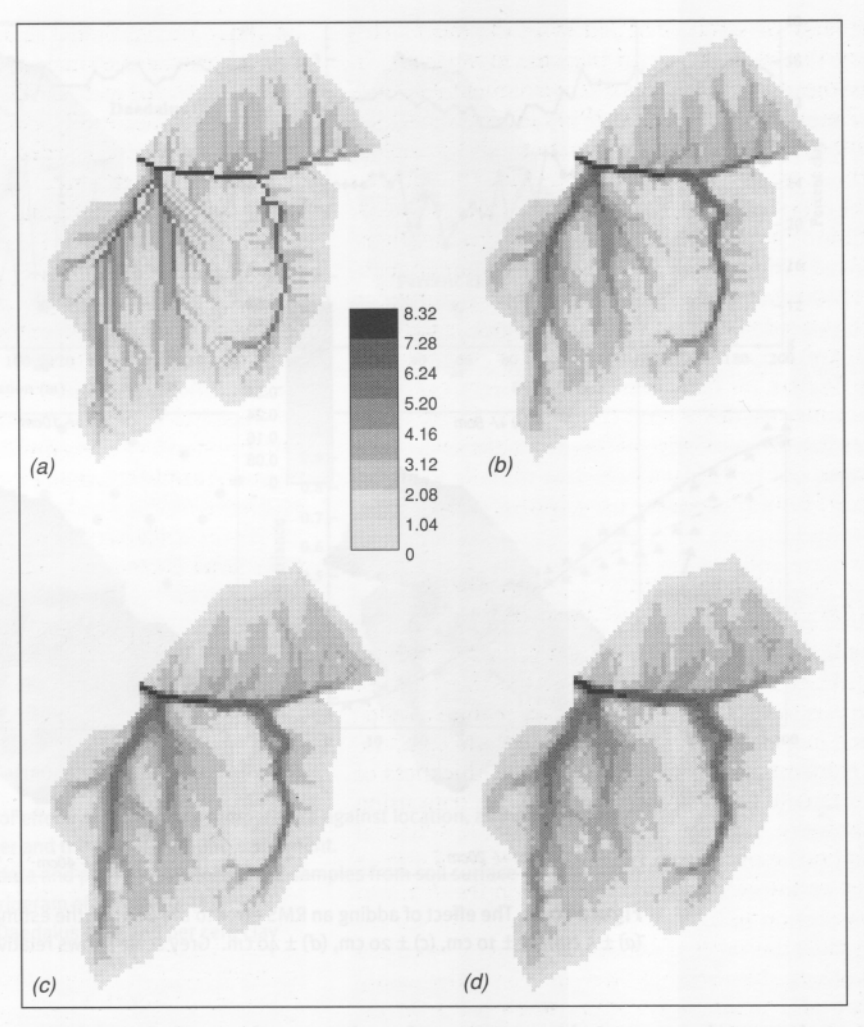


Figure 10.3. The variation in the cumulative upstream elements as a result of different levels of RMS error on the DEM: (*a*) deterministic solution; (*b*) with ± 5 cm error; (*c*) with ± 10 cm error; (*d*) with ± 20 cm error

the direction of the most probable drainage in areas of low relief.

Such information has major implications for the calibration and validation of numerical surface flow models (Desmet 1997). If small relative errors strongly influence the location of computed stream channels then calibrating a model for these channels alone may optimize it for only one out of a large range of possible alternatives. For validating a model, the same holds. If validation measurements are made in locations where the probability is large that a stream flow of magnitude *F* corresponds to a given upstream contributing area, then the validation will be robust. If validation measurements are made in areas where the probabilities are lower, then the chance that the predictions and the validation measurements will converge is naturally smaller. This explains why in hydrological modelling a simple validation based on stream flow at the outlet of the catchment will be robust, while validation measurements made at randomly chosen locations within the catchment may be subject to considerable error.

Clearly modelling using single numbers instead of probability distributions can produce errors of unknown magnitude, particularly if the relative errors are large and it is not known if the single numbers in the database represent modal or average conditions or

some large or small deviation from the mean. This is why the latest trends in *stochastical modelling* in hydrology and other areas of environmental science (De Roo *et al.* 1992; Fisher 1991, Journel 1996) are towards the use of Monte Carlo methods for exploring the complete range of outputs that a model can generate as a result of the probability distributions of both the input data and the model parameters (those numbers in a model used to configure it for application in a given area: the coefficients in a regression model are simple examples). Experience is demonstrating that the methods of conditional simulation aided by 'soft' information about geology, soil, land use, or other controlling variables provide the best inputs for Monte Carlo simulation (Bierkens 1994, Bierkens and Burrough 1993*a*, 1993*b* Gómez-Hérnandez and Srivastava 1990, Journel 1996).

ERROR PROPAGATION THEORY (2): ANALYTICAL APPROACHES TO ERROR PROPAGATION

Though Monte Carlo methods of error analysis are straightforward and can be adapted to many kinds of numerical modelling, even today they require considerable computing resources. For example, to compute the errors associated with mapping the thickness of a sub-surface rocky layer from borehole data using Monte Carlo methods (with or without conditional simulation) requires 100 simulations of the depth to the top of the layer, 100 simulations of the depth to the bottom of the layer, and 10 000 computations of all possible pair-wise differences. It would be useful to have simpler methods that are computationally efficient and faster. Fortunately for many numerical models used in GIS to compute new attributes from existing properties of entities or cells this can be achieved using the standard statistical theory of error propagation (e.g. see Parrat 1961, Taylor 1982).

Numerical operations on single entities or grid cells are known as *point analysis* because no spatial interactions like buffering, topological, or windowing operations are involved. The problem is then for each entity or cell to estimate the error in the output value U as a function of the errors in the input values A_i when the transformation algorithm

$$U = f(A_i) \qquad 10.1$$

includes only arithmetical relationships (+, −, *, ÷, raising to powers, exponentiation, etc.).

Consider the situation in which the value of an attribute on map A_i is not exact but has an associated error term δ so that the value of the attribute cannot be better known than $A \pm \delta a_i$. A_i could be the value of readily available water in a soil mapping unit that is assumed to be statistically homogeneous. We wish to combine the readily available soil water with an estimate of irrigation effectiveness A_j, with an error $A_j \pm \delta a_j$. If the attributes A_i and A_j are statistically independent, and if δa_i and δa_j are each of the order of 20 per cent, it can be shown that the error of total available water δu in the computation of $U = (A_i + A_j)$ is of the order of 28 per cent. For cartographic overlay operations involving more than two steps, the increase in error can be explosive.

Box 10.1 presents the partial differential equations for the simple theory of error propagation. Using these equations we can examine how errors propagate through simple bivariate models with $a_1 = 10 \pm 1$ and $a_2 = 8 \pm 1$.

Sum or difference—no correlation. Let $u = a_i \pm a_j \pm \ldots$, then $\delta u/\delta a_1 = 1$, $\delta u/\delta a_2 = \pm 1$.
By equation

$$(10.4.2)\ S_u = \sqrt{(Sa_1^2 + Sa_2^2)} \qquad 10.2$$

so

$$u = 10 + 8 = 18$$

and

$$S_u = \sqrt{(1 + 1)} = 1.414$$

The absolute error of u is greater than either a_1 or a_2, but in the case of addition, the relative error (1.414/18 = 8 per cent) is lower than for the original variates (10 and 12.5 per cent). For subtraction, the absolute error S_u is the same, but the relative error is now much greater being (1.414/2 = 70 per cent). Whereas addition of two random numbers, and hence of two maps, can be thought of as a benign operation with respect to error propagation, subtraction can lead to explosive increases in relative errors, particularly when a_1 and a_2 are similar in value.

When a_2 is a constant, i.e. $u = a_1 +$ constant, there is no difference in the variance of u and a_1. Adding or subtracting constants has no deleterious effect on errors.

Addition of correlated variables. When the variables $a_1, a_2, \ldots$ are correlated, the term given in equation (10.4.4) must be included in the computation of the error of u. Let $u = a_1 + a_2$, in which $\mathrm{r}a_1a_2$ expresses the correlation ($-1 \le \mathrm{r} \le 1$) between a_1 and a_2.

$$Su = \sqrt{\{Sa_1^2 + Sa_2^2 + 2\,Sa_1.Sa_2.\mathrm{r}a_1a_2\}} \qquad 10.3$$

and

$$Su = \sqrt{\{1 + 1 + 2 \cdot 1 \cdot 1 \cdot \mathrm{r}a_1a_2\}}$$

BOX 10.1. SIMPLE THEORY OF ERROR PROPAGATION

Considering only random, independent errors, for a relationship

$$u = f(a_1, a_2, a_3, \dots a_j) \qquad \text{B10.1.1}$$

in which the a_j's are all independent, *Su*, the standard deviation of *u* is given by

$$Su = \left[\sum_{i=1}^{j} (\delta u/\delta a_i)^2 \cdot Sa_i^2 \right]^{1/2} \qquad \text{B10.1.2}$$

and the standard error of *u*, *SEu*, is given by

$$SEu = \left[\sum_{i=1}^{j} (\delta u/\delta a_i)^2 \cdot SEa_i^2 \right]^{1/2} \qquad \text{B10.1.3}$$

where SEa_i is the standard error of a_i.

These formulae hold when there is no correlation between the x_i's. When they are correlated an extra term must be added to express the increase in error in *u* due to correlation. This term is:

$$\left[\sum_{i=1}^{I} \sum_{j=1}^{J} \{\delta u/\delta a_i \cdot \delta u/\delta a_j \cdot Sa_i \cdot Sa_j \cdot r_{ij} \right] \qquad \text{B10.1.4}$$

If a_i and a_j are 100 per cent positively correlated, the error in *u* can be as much as, but not more than the sum of the errors of a_i and a_j. If a_i and a_j are negatively correlated, the error in *u*, *Su*, could be less than if a_i and a_j were independent.

Product or quotient—no correlation. Let

$$u = a_1^c \cdot a_2^d \qquad 10.4$$

where *c* and *d* are assumed exact constants. Then,

$$\delta u/\delta a_1 = c a_1^{(c-1)} \cdot a_2^d \quad \text{and}$$
$$\delta u/\delta a_2 = d a_1^c \cdot a_2^{(d-1)}$$

so by equation (B10.1.4)

$$Su = \sqrt{\{c^2 \cdot a_1^{2(c-1)} \cdot a_2^{2d} \cdot Sa_1^2 + d^2 \cdot a_1 2c \cdot a_2^{2(d-1)} \cdot Sa_2^2\}} \qquad 10.5$$

Therefore if

$$u = a_1 \cdot a_2,$$

then

$$u = 8 * 10 = 80$$

and

$$Su = \sqrt{\{a_2^2 \cdot Sa_1^2 + a_1^2 \cdot Sa_2^2\}}$$
$$= \sqrt{\{64 * 1 + 100 * 1\}}$$
$$= \sqrt{164}$$
$$= 12.8$$

Multiplication not only raises the absolute error, but also the absolute error, in this case to 12.8/80 = 16%.

When a_j is a constant, *c*, i.e. $u = a_1 \cdot c$, the error propagation reduces to:

$$Su = \sqrt{\{c^2 \cdot Sa_1^2\}} \qquad 10.6$$

Raising to powers

$$\text{For } u = Ca_1^c \qquad 10.7$$

where *C* and *c* are constants, note that a_1 is perfectly correlated with itself so that the error of *u*, *Su* is given by

$$Su = \sqrt{\{C^2 \cdot c^2 \cdot a^{2(c-1)} \cdot Sa_1^2\}} \qquad 10.8$$

For $a = 10 \pm 1$ in the expression $u = a_1^2$

$$u = 10^2 = 100$$

and

$$Su = \sqrt{\{(2a_1)^2 \cdot Sa_1^2\}}$$
$$= \sqrt{\{20^2 \cdot 1\}} = \sqrt{400}$$
$$= 20$$

Not only has the absolute error increased, but the relative error (= 20/100 = 20%) has also doubled.

Logarithmic and other relations.

Let $u = C \ln a_i$ 10.9

then

$$\delta u / \delta a_i = C / a_i$$

so

$$Su = \sqrt{\{(C^2/a_i^2) \cdot Sa_i^2\}} = C \cdot Sa_i/a_i \qquad 10.10$$

Equation (10.24) shows that increase or decrease in error depends solely on the ratio of $C : a_i$.

If

$$u = C \sin a_i$$

then

$$Su = C \cdot Sa_i \cdot \cos a_i \qquad 10.11$$

where Sa_i and a_i are in radians.

SIMPLE EXAMPLES OF ERROR PROPAGATION

Estimating error propagation in net returns for wheat A farmer may wish to estimate the uncertainty associated with his net returns from his wheat fields knowing that the yields and the costs of management and harvesting vary spatially over the farm, and that there is also an uncertainty in the price he will receive. For each field on a tonne per hectare basis he wishes to evaluate the errors in his predictions of

$$\text{Net value}(N) = \text{Yield}(Y) \times \text{price}(P) - \text{costs}(C) \qquad 10.12$$

For each field let Y be $6 \pm 2\text{t} \cdot \text{ha}^{-1}$, P 100 ± 10 currency units per tonne, and C 40 ± 20 currency units per field. The costs per field are assumed to be the same because the farmer has not recorded the actual costs on a field basis, but only for the whole farm.

The gross value

$$G = Y * P = 6 * 100 = 600 \text{ currency units} \qquad 10.13$$

The uncertainty in the gross value is:

$$S_G = {}^1\{P^2 \cdot S_Y^2 + Y^2 \cdot S_P^2\} = {}^1\{10000 * 4 + 36 \quad 100\} = 116.62 \text{ currency units} \qquad 10.14$$

The net value is

$$N = G - C = 600 - 40 = 560 \text{ currency units} \qquad 10.15$$

and the uncertainty is

$$S_N = {}^1\{S_G^2 + S_C^2\} = {}^1\{13600 + 400\} = 118.32 \text{ currency units} \qquad 10.16$$

The error in the Universal Soil Loss Equation Consider the effects of error propagation in one of the modelling exercises discussed in Chapter 7, namely the simulation of soil erosion in Kisii, Kenya. The aim of this study was to estimate the amount of erosion that might occur in a given area under a land use scenario of smallholder maize, grown over a period of forty years. The amount of erosion was estimated using the Universal Soil Loss Equation (USLE—Wischmeier and Smith 1978):

$$A = R * K * L * S * C * P \qquad 10.17$$

where A is the annual soil loss in tonnes ha^{-1}, R is the erosivity of the rainfall, K is the erodibility of the soil, L is the slope length in metres, S the slope in percent, C is the cultivation parameter, and P the protection parameter.

The source data were limited to rainfall studies and conventional choropleth maps displaying only the 'representative' values of the soil and climatic variables. Moreover, the R, L, and S factors were themselves derived through the use of several regression formulae.

The R factor. The value of R was estimated by using the FAO formula:

$$R = 0.11abc + 66 \qquad 10.18$$

where a is the average annual precipitation in cm, b is the maximum day precipitation occurring once in 2 years in cm, and c is the maximum total precipitation of a shower of one year occurring once in 2 years, also in cm. The best information available for the study area (Wielemaker and Boxem 1982) suggested that reasonable values for a, b, and c were:

$$a = 172.5 \pm 20 \text{ cm}, \; b = 5.41 \pm 1.1 \text{ cm}, \; c = 2.25 \pm 0.5 \text{ cm}.$$

The estimates of the standard errors are large because of the limited data for the area.

By equation 10.18, $R = 297$ cm and by Box 10.1, the standard error of R is $\pm$ 72 cm per year.

The K factor. For the more erodible soils, evidence from outside the study area suggested that a K value of 0.1 for an eroded soil over laterite was probable.

Given the often large variability of soil properties (a CV of 50 per cent is common) a map unit standard deviation of ± 0.05 is not unreasonable.

The L factor. The slope length factor L was calculated using Wischmeier and Smith's (1978) formula of

$$L = (l/221)^{1/2} \qquad 10.19$$

where l is the slope length in metres. Considering a typical, long slope in the area, a length of 100 m ±20 m is reasonable; this converts to a L value of 2.13 ± 0.045.

The S factor. The slope gradient factor S was computed using the parabolic regression formula

$$S = 0.0065s^2 + 0.0454s + 0.065 \qquad 10.20$$

where s is the slope expressed as a percentage (Smith and Wischmeier 1957). For a slope of 10 ± 2 per cent (an optimistic estimate of the standard deviation of the classes on the slope map) equations 6.57 and 6.47 yield a value for S of 1.169 ± 0.122.

The C factor. The estimated value of the crop management factor C was 0.63. Given the uncertainty of estimating C for a crop that does not cover the soil surface for the whole year, and that Rose (1975) found that a C-value for maize varies between 0.4 and 0.9, the error in C was estimated at ± 0.15.

The P factor. The erosion control practice factor, P, was estimated at 0.5 ± 0.1.

To summarize, the values of the factors of the USLE and their estimated errors are:

$$R = 297 \pm 72$$
$$K = 0.1 \pm 0.05$$
$$L = 2.13 \pm 0.045$$
$$S = 1.169 \pm 0.122$$
$$C = 0.63 \pm 0.15$$
$$P = 0.5 \pm 0.1$$

which yields an annual soil loss rate of 23 ± 14.8 tonnes per hectare per year, corresponding to a soil surface lowering of 0.23 ± 0.15 cm per year, or 9 ± 6 cm in 40 years. In other words, the variation in predicted soil loss, given the integrity of the model, was that 95 per cent of cells having the climate/slope/soil regime specified here would have a soil loss ranging between 3 and 21 cm. The 95 per cent confidence limits on A in tonnes $ha^{-1}y^{-1}$ are $-7.0 < 23.0 < 53$ when $r = 0$ and $-17.0 < 23 < 63$ when $r = 0.25$, which shows the deleterious effect that a small, positive intercorrelation can have on the error band.

Recommended strategies for arithmetic algorithms

The theory and examples provide the following rules to reduce error propagation:

1. **Avoid intercorrelated variables**
2. **Add where possible**
3. **If you cannot add, multiply or divide**
4. **Avoid as far as possible taking differences or raising variables to powers.**

These rules of thumb have considerable implications for the quality of the results obtained by feeding models of environmental process by data from a GIS. They demonstrate that in order to reduce the errors in the output of point models it is important to reduce the errors of the model inputs. In spatial modelling, errors in inputs can be reduced by:

- **(a) using optimal interpolation techniques**
- **(b) by using the appropriate sampling density**
- **(c) removing or checking for outliers, subgroups, systematic bias, etc.**
- **(d) adopting appropriate classifications**
- **(e) using sensible models**
- **(f) improving model calibration by reducing errors in model parameters**

Error propagation tools—ADAM

The methods given above can easily be worked out on a pocket calculator for single equations with only two or three variables but the computations become tedious when they must be repeated for a large number of cells or entities in a database, or when numerical models involve larger numbers of attributes, including correlation. Recently Gerard Heuvelink and Cees Wesseling (both then at Utrecht University) wrote a computer program called ADAM which can trace errors through complex numerical models in 'point mode' that operate on the attributes of entities or on multiple raster overlays (Heuvelink 1993, Heuvelink *et al.* 1989).

Consider a multiple regression model in which the a_i values are raster maps of input attributes and the b_i values are the coefficients.

$$u = b_0 + b_1a_1 + b_2a_2 + \ldots + b_na_n \qquad 10.21$$

Contributions to the error in output map u come from the errors associated with the model coefficients b_i and from the errors associated with the spatial variation and measurement errors (nugget) of the a_i input attributes.

Errors associated with the model coefficients It is not always easy to determine the errors associated with model coefficients, but for multiple regression models the errors are given by the standard errors of the coefficients b_i and their correlations. All these terms should be provided by a good statistical package.

Errors accruing from spatial variation These can be derived in several ways, including intuition, and sampling combined with conventional or spatial statistics. Intuitive estimates of the standard deviation of attribute values can provide error estimates if there are no hard data. Alternatively, sampling within spatial entities like soil or vegetation polygons can provide useful information on the estimated standard deviations of the attributes and their correlations. If there are sufficient data to interpolate to raster overlays using methods like kriging or conditional simulation then the standard deviation per cell is a possible source of the required uncertainties. Both co-kriging and co-conditional simulation (Deutsch and Journel 1992) can be used to generate correlation surfaces, though if the data are reasonably evenly distributed over the area, the simple correlation coefficient between the different attributes r_{ij} will suffice.

The balance of errors The errors δu in the output maps of u accrue from all sources, be they coefficients or variables and it is useful to know which of these is the major source of uncertainty. This knowledge can be used to improve either the model or the data collection if deemed necessary. Comparing error analyses for different models (e.g. a single term regression versus a multiple term regression) would allow users to decide whether collecting data on a second attribute was worth the time and money. Comparing error contributions between model and data provides a useful basis for optimizing survey effort between model calibration and mapping.

The theory behind ADAM. Equation (10.35) above, can be generalized to:

$$y = g(z_1, z_2, \ldots, z_n) \qquad 10.22$$

where g is a continuously differentiable function from R^n into R. The arguments z_i of g may consist of both the input attributes A_i relating to an entity or grid cell and the model coefficients b_i. The problem is to determine the error in the y caused by errors in the z_i. To do this we assume that y, z_i are realizations of random variables Y, Z_i. For notational convenience we define vectors $\mu = [\mu_1, \mu_2, \ldots, \mu_n]^T$ and $Z = [Z_1, Z_2, \ldots, Z_n]^T$.

The simplest way to compute the quantities of interest, namely the mean and variance of Y, is to use a Taylor Series expansion of g around μ, neglecting higher-order terms (see Heuvelink *et al.* 1989 for details).

Note that when the correlation between the arguments Z_i is zero, the variance of Y is

$$\sigma^2 = \sum_{i=1}^{n} \{\sigma_i^2 (\delta g / \delta z_i \cdot (\mu))^2\} \qquad 10.23$$

This means that the variance of Y is the sum of parts, each attributable to an input Z_i. This partitioning property allows one to analyse the contribution of each input, be it regression coefficient or variable, to the final error.

ADAM uses this technique to apply the error propagation to each entity, which in the case of a raster map is the grid of pixels, but in other implementations could be polygons or other entities in vector mode (Wesseling and Heuvelink 1991). The regression model coefficients and their errors can be computed from sample data and the same point samples can be used to create raster input maps of cell values and their errors. ADAM computes the model output and the errors as maps, and also maps of the spatial variation of the error contribution from all coefficients and input variables (Figure 10.4).

AN EXAMPLE OF COMPUTING ERROR PROPAGATION IN REGRESSION MODELLING WITH ADAM

Chapters 5 and 6 presented a range of numerical methods for predicting the concentrations of heavy metals polluting the part of the flood plain of the River Maas in the south of the Netherlands. The options included regression modelling based on a multiple regression analysis of heavy metal concentration, dis-

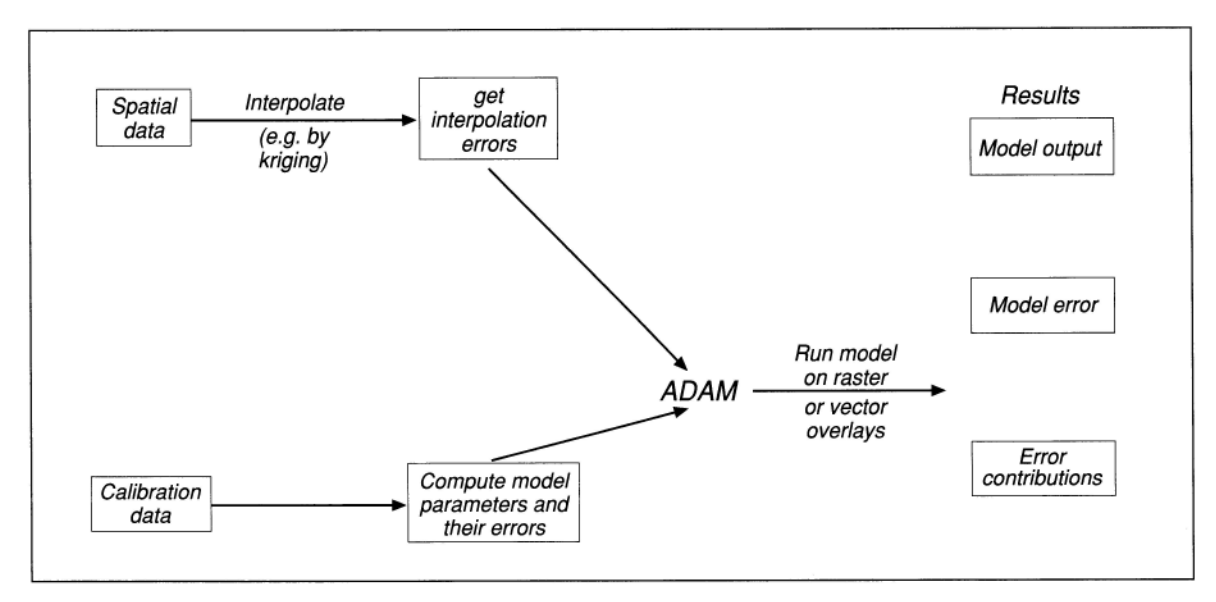


Figure 10.4. Flow chart of operations when using the ADAM procedure of follow the accumulation of errors in regression models in GIS

tance from the river, and relative elevation and direct geostatistical interpolation by ordinary kriging or co-kriging. Though several authors (Lam 1983, Laslett *et al.* 1987, Leenaers *et al.* 1990, Englund 1990) have compared the performance of different interpolation techniques this has rarely been done in a cost-benefit context. We use the ADAM error propagation tool and independent validation samples to compare different interpolation techniques in order better to understand the circumstances under which approach gives not the best interpolation, but the best value for money. In this example we use the same basic data set as in Chapters 5 and 6 (but from a slightly larger area). This time the complete data set of 155 soil samples is split into a validation set of 53 samples drawn at random and 102 samples which are used for interpolation. Statistical comparisons of the two sub-data sets suggested that they were not significantly different samples of the study area and that the validation data would provide a good test of the predictions.

Preliminary analysis (Chapter 5) showed that the zinc content was strongly correlated with the attributes 'distance to the river' and 'elevation'. Because both zinc content and 'distance to river' were strongly log-normally distributed, they were transformed to natural logarithms and all further statistical analysis uses the log-normally transformed data. Correlation analysis showed strong negative correlations between ln(zinc) and ln(distance) [–0.768], between ln(zinc) and relative elevation [–0.624] and a positive correlation between ln(distance) and relative elevation [0.481]. The multiple regression

$$ln(zinc) = B_0 + B_1 \cdot ln(distance) + B_2 \cdot elevation + \varepsilon \qquad 10.24$$

was fitted with an $R^2 = 0.74$, when the regression was based on 102 samples, so this appears to be a reasonable model of the dependence of zinc concentration on the other two attributes.

Procedure We compare the following *options* for mapping the zinc content of the floodplain soils:

(*a*) Use choropleth maps of soil type or flood frequency and compute means and standard deviations per mapping unit, assuming within unit variation is given by the mean and variance of the included sample sites.

(*b*) Map the zinc content by applying the bivariate linear regression equation (10.24) to maps of distance to the river and relative elevation that have been obtained by ordinary kriging. The kriging error maps for the distance and elevation provide the sources of the errors for the attributes.

(*c*) Interpolate the zinc content directly using ordinary kriging

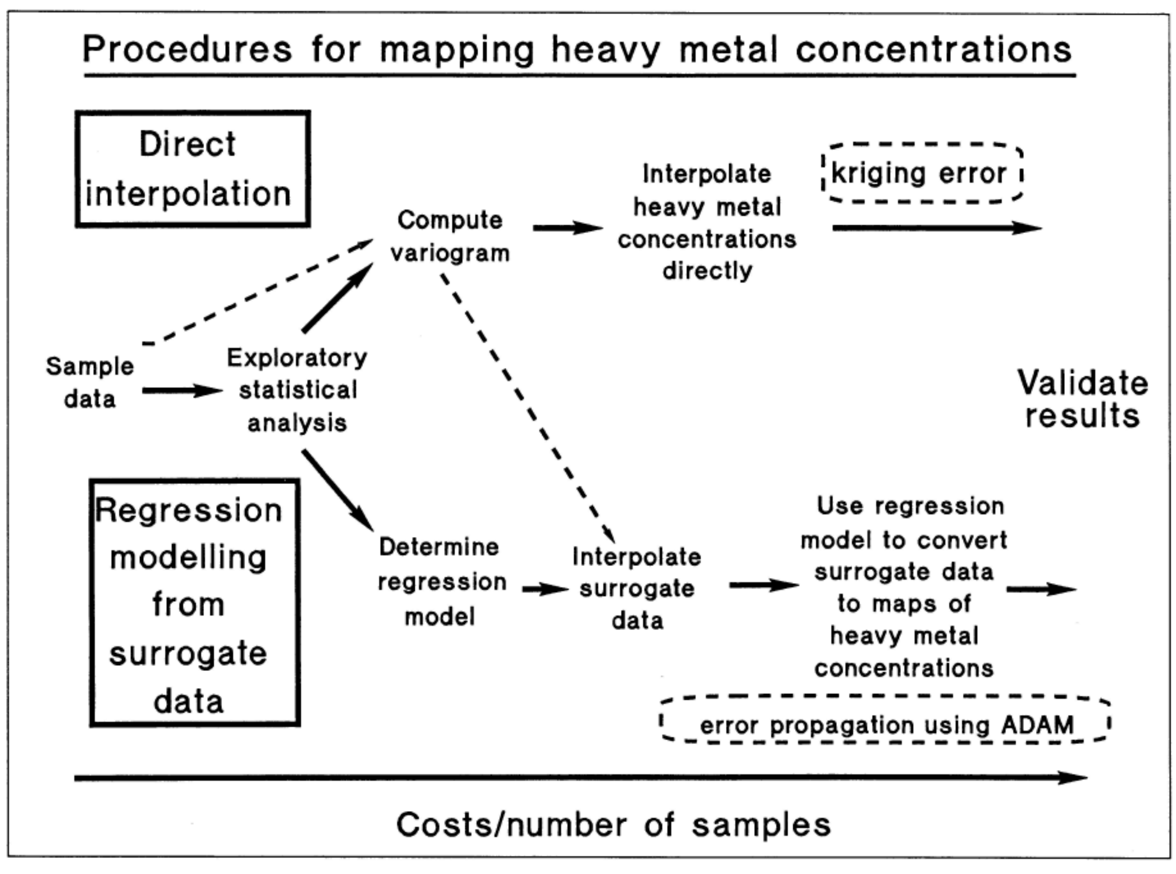


Figure 10.5. The procedures for mapping heavy metals in floodplain soils by direct interpolation or regression modelling

Figure 10.5 summarizes the procedures for options (*b*) and (*c*).

The *costs* of each method depend on the number of observations and types of samples used. Distance measurements are very cheap because they can be read off a map, or generated by a buffer command; elevation data can be read cheaply off a detailed map or can be recorded more accurately in the field, where the major cost is the cost of labour. Determining the levels of heavy metals in soil samples is expensive because samples must be located and collected by hand, and laboratory analyses are expensive. Therefore it is possible that under some circumstances better results might be obtained by using fewer direct measurements of the zinc content of the soil and more of the cheaply measurable site attributes.

To see how the quality of results depend on the *numbers of samples* the data set of 102 samples was used to create several smaller data sets, namely two sets of 51 samples, two sets of 27 samples, and two sets of 14 samples. The costs of collecting and processing the data for each data set were computed using commercial rates for the attributes in question.

Each method was applied to all data sets to predict values for 20 × 20 m grid cells; results for the data sets of the same size were pooled. Error propagation using ADAM, kriging standard errors, and independent validation by the 53 test samples were used as criteria of success.

For regression modelling it was assumed that the input maps were always those interpolated from the 102 sites, and the variograms of both attributes are given in Figure 10.6. All data were interpolated by block kriging to a 20 × 20 m grid (Figure 10.7). The coefficients and correlations of the regression model were determined for each sub-data set separately using 14, 27, 51, or 102 zinc data respectively to provide a range of models with different goodness of fit.

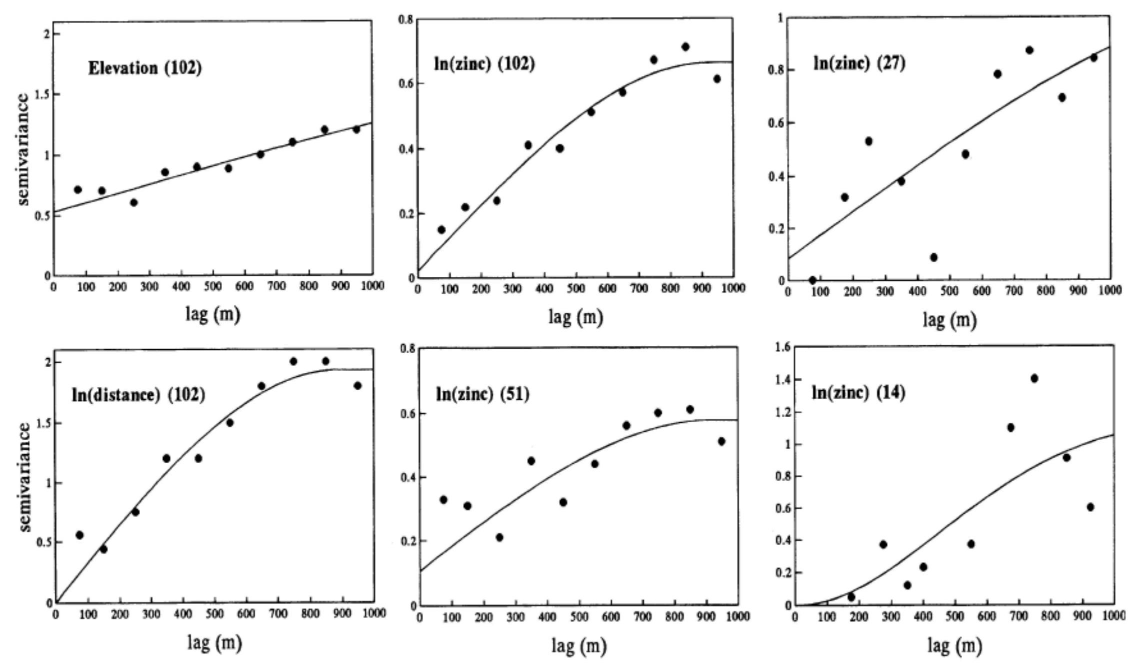


Figure 10.6. Variograms of key attributes in the cost/benefit analysis of interpolation. Top left: floodplain elevation using 102 samples; top centre: ln(zinc) using 102 samples; top right: ln(zinc) using 27 samples; bottom left: ln(distance to river) using 102 samples; bottom centre: ln(zinc) using 51 samples; bottom right: ln(zinc) using 14 samples

ADAM used each model to compute a map of zinc corresponding to the number of zinc samples used; these maps were validated using the 53 test data.

Box 10.2 shows the control file for the linear regression model with the model coefficients, their standard errors and correlations, and the model results and error surfaces. ADAM recommended that this model could be approximated by a second order Taylor series. Box 10.3 shows the results of automatically translating the control file into a set of commands in the PCRaster raster modelling system (Wesseling *et al.* 1996); this shows the difficulties of trying to write error propagation routines by hand.

The results of computing the ln(zinc) surfaces by regression from example sets of 27 and 102 data points, respectively, are shown in Figures 10.8*b* and 10.8*d*.

The balance of errors ADAM apportioned the error of the inputs according to contributions from the distance to the river, the elevation, and the model coefficients and presented them as continuous maps (Figure 10.9). These maps show that at grid cells coinciding with sample locations, the model is the largest source of error, sometimes reaching 65 per cent. The contributions from 'distance to river' are greatest where this attribute has not been measured, such as at the eastern boundary of the area. The contributions from 'elevation' also peak at the sample points, possibly because this property has a low overall variation and at the sample sites the local deviations are significant in terms of the small overall range of relief over the site and its large nugget compared to the error in the logarithm of distance from the river.

The errors propagated by the regression models based on example sets of 27 and 102 data points are given in Figures 10.8*f* and 10.8*h*.

Direct interpolation For mapping the ln(zinc) directly, variograms were computed for each subset of the data and used for interpolating from the same data set. No variograms could be modelled for the data sets with 14 sites: variograms based on 27 data were poorly defined—see Figure 10.6.

Figures 10.8*a* and 10.8*e*, and Figures 10.8*c* and 10.8*g* show the maps and error surfaces obtained by ordinary kriging for example data sets of 27 and 102 samples, respectively. Figure 10.8 shows that ordinary kriging using 102 samples gives the best combination

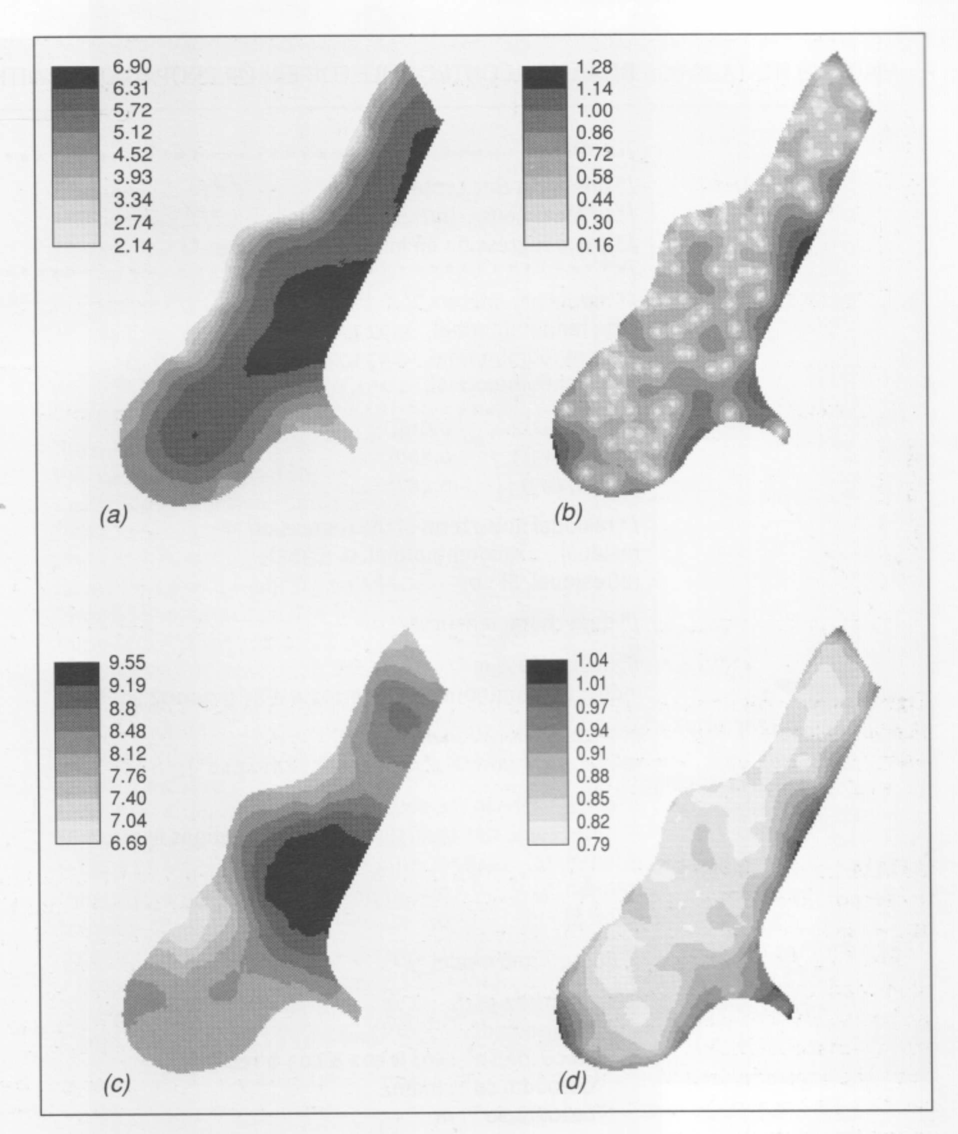


Figure 10.7. Interpolated surfaces (predictions and errors) used as inputs for the regression modelling of ln(zinc): (*a*) predictions of ln(distance to river); (*b*) prediction variance for ln(distance to river); (*c*) predictions of elevation (m); (*d*) prediction variance for elevation

of spatial resolution and low predicted errors, but kriging using only 27 sites is clearly unsatisfactory and worse than either regression model. The differences between the two regression model results are small, implying that increasing the number of zinc samples from 27 to 102 has not much improved the maps.

Similar studies were carried out using the maps of soil types and flood frequency. All maps were validated by computing the validation standard deviation $VSD = \sqrt{[(\hat{z}(x_i) - z(x_i))^2/53]}$ using the 53 data points set aside for this purpose. For each map produced by interpolation or modelling, the mean *prediction standard deviation* (*PSD*) was calculated from the error surfaces shown in Figure 10.8*e-h*.

Costs and benefits Figure 10.10 shows how the mean prediction standard deviation (*PSD*) and the validation error (*VSD*) vary with survey costs for the flood frequency map, the regression modelling, and the ordinary kriging of zinc. Figure 10.10 shows

BOX 10.2. CONTROL FILE FOR ERROR PROPAGATION WITH ADAM

```
/*********************************************/
/* predict ln zinc content                  */
/* of Maas soils    (lnzkp)                 */
/* linear regression on lnDM & RA-102 sites */
/*********************************************/
/* model parameters */
co1 : random(normal,  9.973, 0.299);
co2 : random(normal, -0.333, 0.333);
co3 : random(normal, -0.291, 0.041);

ro(co1, co2) :      -0.014;
ro(co1, co3) :      -0.860;
ro(co2, co3) :      -0.487;

/* residual noise term of the regression */
residual   : random(normal, 0, 0.365);
ro(residual, $) : 0;

/* data characteristics*/

/* ln dist Mass m */
lndm : random(normal, "lndm102.est", "lndm102.sd");

/* relative elevation m */
ra : random( normal, "ra102.est", "ra102.sd");

/* correlations in the data */
/* note: could use spatially varying correlations in errors */
ro(lndm, ra) : 0.4870;
ro(lndm, $)  :       0;
ro(ra, $)    :       0;

/* output from model */

calc("lznm102.est",
     "lznm102.sd",
     "zmod102.sd" : co1 & co2 & co3 & residual,
     "da102dm.sd" : lndm,
     "da102ra.sd" : ra
   ) : co1 + co2*lndm + co3*ra + residual;
```

clearly that when only a few samples (<20) can be afforded the simple flood frequency maps score best (and the predicted standard errors are similar to the validation errors), but using more data to improve the precision of the means and standard deviations of the mapping units does not improve spatial prediction. Clearly, as shown in Chapter 6, there is much spatial variation within the flood frequency classes which is not detected by computing simple area-average statistics.

Once more than 20 samples are used the empirical regression model performs better than the flood frequency map (the validation results are much better than predicted!) but once the spatial trends described by the regression have been modelled correctly, little further improvement in either prediction or validation is obtained by using more than 35 sites to estimate the regression model.

With insufficient data (<50 samples in this case) useful variograms cannot be modelled; interpolation

BOX 10.3. COMMAND FILE FOR COMPUTING ERROR PROPAGATION BY ADAM

```
/****************************************/
/*                            */
/* This script is generated by ADAM, */
/*   an error propagation tool      */
/*   developed at the               */
/*   University of Utrecht          */
/*   by C. G. Wesseling &            */
/*       G. B. M. Heuvelink          */
/*                            */
/****************************************/
tmp0.$$$ = -0.291*ra102.sd;
tmp1.$$$ = sqr(tmp0.$$$);
da102ra.sd = cont(sqrt(tmp1.$$$));
tmp2.$$$ = -0.333*lndm102.sd;
tmp3.$$$ = sqr(tmp2.$$$);
da102dm.sd = cont(sqrt(tmp3.$$$));
tmp4.$$$ = lndm102.est*0.033;
tmp5.$$$ = ra102.est*0.041
tmp6.$$$ = (((0.299*tmp4.$$$)*-0.014 + ((0.299*tmp5.$$$)*-0.86)) +
((tmp4.$$$*tmp5.$$$)*-0.487;
tmp5.$$$ = ((sqr(0.299) + sqr(tmp4.$$$)) + sqr(tmp5.$$$)) + sqr(0.365);
zmod102.sd = cont(sqrt(((2.0*tmp6.$$$) + tmp5.$$$)));
lznm102.est = cont((9.973 + (-0.333*lndm102.est)) + (-0.291*ra102.est));
tmp6.$$$ = tmp6.$$$ + ((tmp0.$$$*tmp2.$$$)*0.487);
tmp0.$$$ = sqr(lndm102.sd)*sqr(0.033);
tmp2.$$$ = (-0.487*(0.033*0.041))*(0.487*(ra102.sd*lndm102.sd));
tmp4.$$$ = sqr(ra102.sd)*sqr(0.041);
lznm102.sd =
cont(sqrt((((2.0*tmp6.$$$) + (tmp1.$$$ + (tmp3.$$$ + tmp5.$$$))) + (0.25*(tmp4.$$$
+ (tmp2.$$$ + (tmp4.$$$ + (tmp2.$$$ + (tmp2.$$$ + (tmp0.$$$ + (tmp2.$$$ +
(tmp0.$$$ + (tmp4.$$$ + (tmp2.$$$ + ((tmp2.$$$ + (tmp2.$$$ + (tmp0.$$$ +
(tmp0.$$$ + tmp2.$$$)))) + tmp4.$$$))))))))))))));
```

from sparse data brings larger interpolation errors which are confirmed by the lower validation errors. Insufficient data points also mean that the interpolation algorithm cannot make predictions for the whole area, as Figures 10.8*a* and *e* show. When more than 50 samples can be afforded, kriging interpolation is superior because there are enough 'hard' data points to resolve spatial variation over distances that cannot be resolved by the global regression model. Here we note in passing that a still more effective approach might be to combine the regression with the kriging in the KT-universal kriging model (Deutsch and Journel 1992) to combine knowledge of the global trend with optimal estimates of local variation (see Chapter 6).

Figure 10.10 shows that generally the validation errors are smaller than predicted errors, which suggests that all the methods are performing better than expected. This is caused by the large interpolation errors in unsampled (built up) parts of the area, but these areas were not included in the validation set. Therefore the validation data sampled a spatially more homogeneous area than the whole survey. If allowance is made for these unsampled areas the mean *PSD* and *VSD* values are much closer.

ln(zinc) (ln (ppm))
7.49
7.15
6.80
6.46
6.11
5.77
5.42
5.08
4.73

(a) (b) (c) (d)

estimation standard deviation (ln(ppm))
0.90
0.82
0.73
0.64
0.55
0.46
0.37
0.28
0.19

(e) (f) (g) (h)

Figure 10.8. The comparison of predictions of ln(zinc) in floodplain soils made by regression and by ordinary point kriging, using different numbers of samples (*a*) kriging using 27 zinc samples for variogram and interpolation; (*b*) regression modelling using 27 zinc samples and ln(distance to river) and elevation based on 102 samples; (*c*) kriging using 102 samples for variogram and interpolation; (*d*) regression modelling using 102 zinc samples and ln(distance to river) and elevation based on 102 samples
Maps *e–h* show the respective error surfaces for maps *a–d*, expressed as standard deviations of ln(zinc)

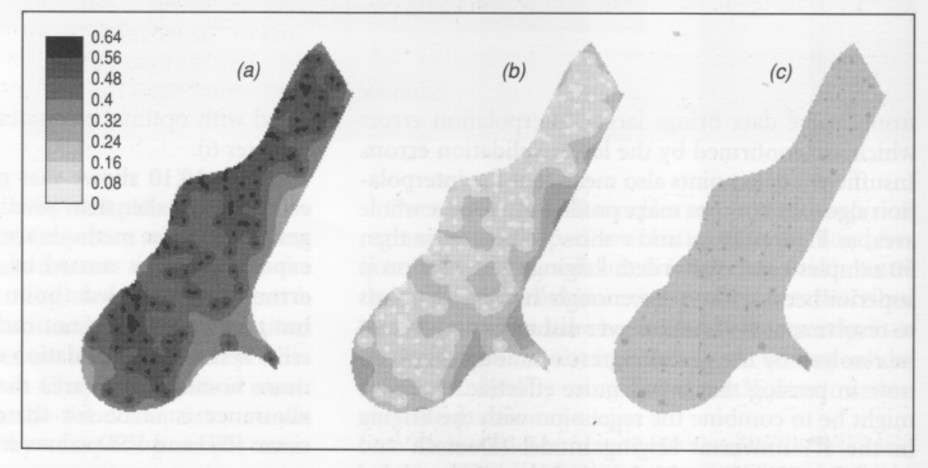


Figure 10.9. The contributions of inputs to the errors in the maps obtained by regression modelling of ln(zinc) from 102 samples (*a*) Contributions of the regression model; (*b*) Contributions of ln(distance to river); (*c*) Contributions of elevation

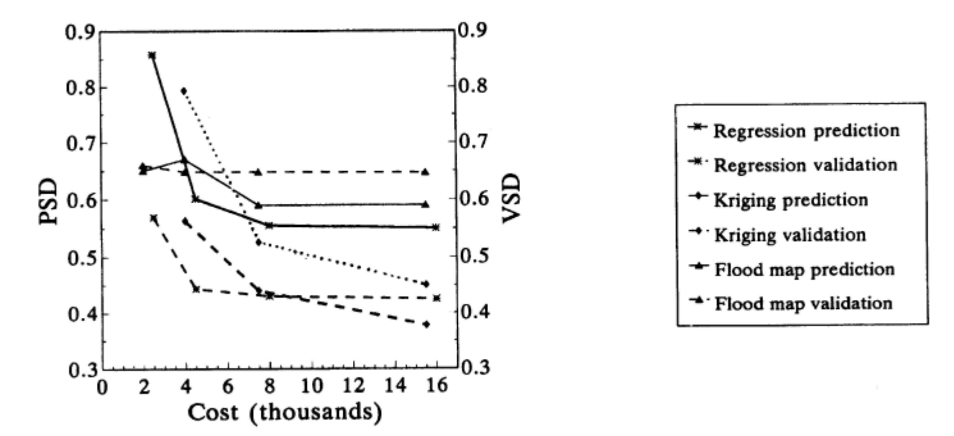


Figure 10.10. Cost-quality comparisons for the different interpolation methods, expressed in terms of their own prediction standard deviations (PSD) and standard deviations obtained from 55 independent, validation data (VSD)

Using the variogram to optimize sampling networks and reduce errors at the cost of reduced spatial resolution

At the beginning of this chapter we demonstrated that sample networks and sample sizes (the support) perform best when they are tuned to the natural scales of variation. This is stated empirically by the *sampling theorem* (which was originally derived from experience in electronic signal processing)—observations must be made at a frequency and resolution (i.e. support) that matches the spatial correlation structure of the phenomenon in question. If sampling frequency and size of sample are ill-chosen then the patterns of variation cannot be resolved and will largely appear as noise.

The lessons to be learned

Though these results are not necessarily general for all rivers and for all kinds of pollution modelling the study demonstrates the value of correctly identifying the spatial correlation structures (patterns) and their relation to physical processes. They demonstrate the advantages of tuning both the interpolation and the sampling density to the spatial correlation structure of the attributes being mapped and show that costs can be saved by using appropriate techniques.

As many natural phenomena vary at many different scales (i.e. have fractal-like behaviour—Burrough 1983*ab*, 1993*a*, Mandelbrot 1982) most sampling networks will pick up some structured information but they may not be providing the best quality data for the investments made. In Chapter 5 we pointed out the use of *bulking* in sampling to remove short-range spatial variation before interpolation, and in Chapter 6 we showed that with block kriging we can make predictions for a block of land larger than the support with a smaller kriging variance. We also showed in Chapter 6 (e.g. 6.17) that because the block kriging variance depends on (*a*) the variogram, (*b*) the block size, and (*c*) the configuration of the data points, it is possible, if the variogram is known, to derive quantitative relations between within-block variance and sample spacing on a regular grid. These relations can be used to choose the best sample spacing for a given *PSDn* for a given budget. McBratney and Webster (1981) wrote a program called *OSSFIM* for this purpose.

Increasing block size, however, reduces prediction errors but the consequence is that one must be prepared to deal with larger blocks of land. If interpolated data are to be used in numerical models then it is sensible to match the sizes of the cells in the interpolated surface to the sizes of cells used in the model, and this has direct consequences for error propagation. The relation between block size and error can also be important when decisions must be made about how land must be treated. For example, if the level of zinc pollution is deemed to be too high, then the soil may have to be removed or treated. Small units (blocks) are desirable because they can be cleaned up for less money than large blocks. If the errors associated with the means of small blocks are too large, however, there is a reasonable chance that blocks that really fall below the acceptable threshold levels must be treated and blocks that are really above that level are missed.

Using the variogram to optimize the zinc sampling net

The comparison of costs and benefits of different methods of interpolation for the zinc data was based on the original sampling pattern, which was the same in all cases (except for sub-sampling). If we had known the variograms of the input data (elevation, distance-to-river, zinc) before the field survey had been carried out, and if we had been prepared to accept interpolated mean values for a block of land side *B* instead of for areas the size of the support we could have computed the best possible sample spacing *for mapping* for any given level of sampling investment.

Figure 10.11 (top) shows the normalized variograms of ln(zinc) and elevation. Clearly, the spatial correlation structure for zinc is stronger than elevation, as shown by the larger nugget for elevation and the greater deviation of estimated semivariance from the model. Figure 10.11 (middle) shows the graphs of sampling spacing versus block size obtained using *OSSFIM* for both attributes—one sees that for a given number of samples, better results can be obtained from ln(zinc) than elevation, but the advantage of the spatial structure of zinc is offset by the cost of measuring it. So if we are going to use maps of elevation to help predict zinc it would be useful to compute the sample spacing on a square grid and the block size combinations that would bring about improvements.

Figure 10.11 (bottom) shows how the normalized kriging variance for these two attributes varies with sample spacing (and hence costs). It demonstrates that in this area, sampling elevation on a 50 m net and computing 100 m block averages yields a relative precision that is as good as sampling zinc on a 150 m net for the same block size. This means that when sufficient zinc samples can be afforded it will be better in this case to opt to interpolate zinc directly, and not to compute from a regression model involving elevation, *even though the elevation is recorded on a net that is three times as dense as the sample net for zinc.* Put another way, increasing sampling density on a regular square grid for zinc brings more benefit per sampling point than when measuring elevation.

Applying this knowledge to the costs and sample numbers used in the comparative study allows as to compare the levels of *PSD* and *VSD* obtained for kriging from the *original sampling net* with the maximum expected kriging standard deviations for a *range of block sizes on regular square grids* having as many data points as were actually used (14, 27, 52, 102). These results are shown in Figure 10.12. They suggest that in this case optimizing the sampling net to a square grid for mapping would not bring great benefits, as compared with the validation of the actual point kriging interpolation. If the variogram was unknown, sampling on sparser grids would also hinder variogram estimation as there would be no close pairs of points to compute semivariance at short lags.

Referring back to Chapter 6, Table 6.2, although we

Geostatistics in error propagation

Geostatistics have an important role to play in reducing errors in numerical modelling with GIS. Geostatistics can provide information on interpolation uncertainty and can be used to provide optimal estimates of attribute values for spatial blocks of given size and known variance. Examination of variograms can show whether different data are spatially compatible. Knowledge of the variogram can be used to optimize the geometry of regular sampling patterns used for interpolation mapping. Geostatistics linked to error propagation by Monte Carlo or analytical methods can be used to follow the propagation of errors through numerical modelling.

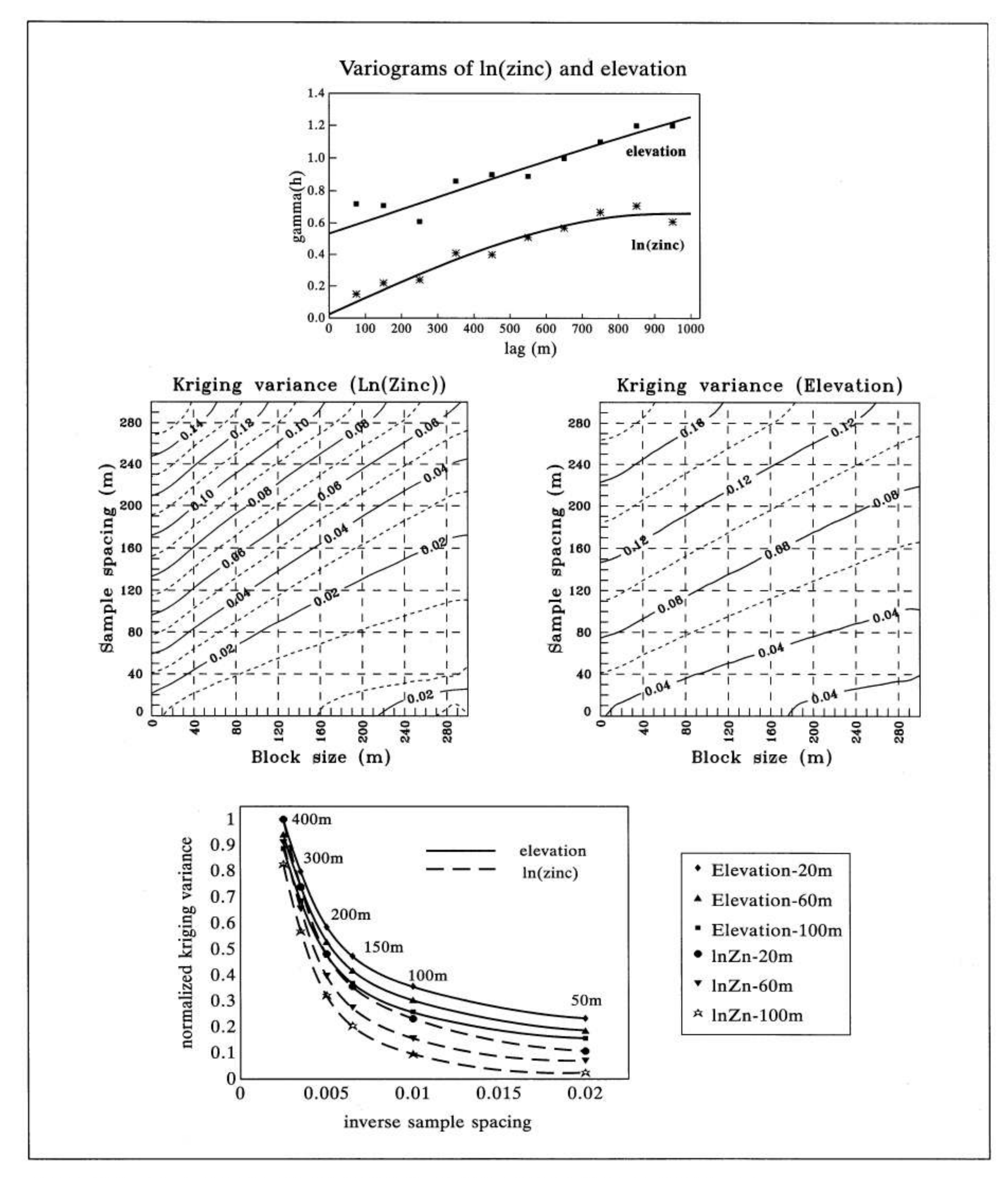


Figure 10.11. The relations between variogram (top), sample spacing and block size (middle), and the relative efficiency of sampling networks on a square grid (bottom) for mapping ln(zinc) and flood plain elevation

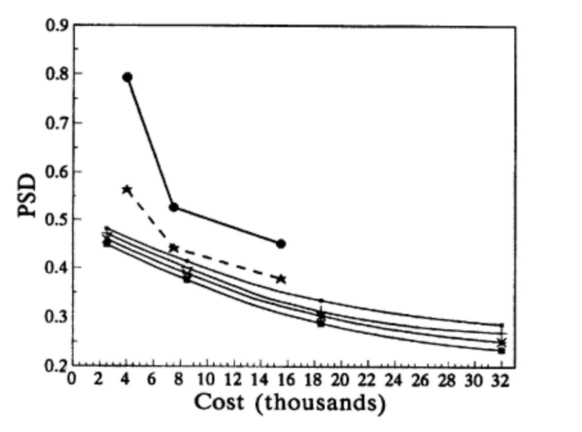


Figure 10.12. Cost-quality comparison of the results of ordinary kriging using the actual sample net with (theoretical) optimized networks

cannot directly compare results computed on a log scale with untransformed data, there is a strong suggestion that the best way to improve mapping in this area is to include 'soft' information, for example by stratifying according to flooding frequency rather than opting for a regular square grid. The reason in this case is that the spatial resolution of the flood frequency map more than makes up for a sparse sampling network. This demonstrates the value of using geostatistics and GIS together. As an exercise, readers might like to work out for themselves whether this is really so.

Intelligent GIS

Using error analysis for optimizing spatial modelling in GIS

This and the previous chapter have presented many aspects of how errors accrue in spatial data and spatial data analysis. We have dealt with several methods for assessing loss of information through errors and how to cope with problems like vector-raster conversion, error propagation in numerical modelling, the matching of correlation structures, and optimizing sample networks. The challenge now is to put these error tracking methods together in such a way that an intelligent GIS can indicate to a user whether a particular spatial analysis problem can be solved to an intrinsic level of quality, and if not, to advise on the steps to be taken to correct matters. An intelligent GIS would include formal rules to help the user choose the best set of procedures and tools to solve the problem within the constraints of data, data quality, cost, and accuracy. A really intelligent GIS would be able to carry out error propagation studies *before* a major data crunching operation to estimate if the methods and data chosen were likely to yield the results intended. It would report to the user where the major sources of error come from and would present him or her with a set of options which would achieve better results.

In spite of pessimism by Dunn *et al.* (1990), we believe that systematic error checking in GIS modelling is now possible. This chapter has demonstrated that there are methods available for following errors through GIS operations, and that once error propagation is understood, the user will be able to select those methods that lead to better results. The options for improvement include:

(*a*) using better methods for spatial interpolation or use numerical models instead of simple logic

(*b*) collect more data and optimize sampling
(*c*) collect different data
(*d*) use better models—match data and models
(*e*) improve the model calibration
(*f*) improve the spatial and/or temporal resolution by matching correlation structures.

Each option would be accompanied by an estimate of the costs so that rational decisions could be taken (Figure 10.13). The system would also be able to indicate situations in which the results were *much better* than expected: in these cases important savings on data collection and processing could be made without serious loss of information.

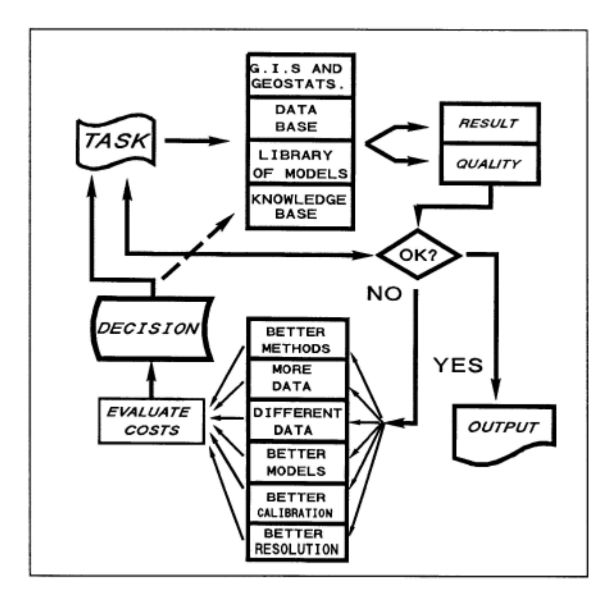


Figure 10.13. An intelligent GIS would be able to advise a user on how best to obtain a result that meets given quality standards

END NOTE

This chapter has demonstrated the need for proper descriptions of conceptual ideas, data collection methodology, and analysis techniques archived in the metadata accompanying all data sources, which must be written in appropriate standards. There is increasing interest interest in the role of data accuracy and errors in GIS (e.g. Guptill and Morrison 1995, Thapa and Bossler 1992), though as yet no general, integrated practical tools for statistical error propagation in GIS exist. Many important components have been developed and are readily available to researchers and several research studies on error propagation are current (Lanter and Veregin 1992, Lodwick *et al.* 1990, Heuvelink 1993).

Questions

1. Discuss the value and the difficulties of getting a good idea of spatial correlation structures when using GIS for environmental modelling.

2. How do you think the errors in the output of any given model will depend (*a*) on the uncertainties in the regression equation, (*b*) the spatial variation in the data? How would you investigate this quantitatively?

3. How would you design a cost-benefit study to relate the prediction errors to (*a*) numbers of model calibration points? (*b*) numbers of sites used for interpolation? (*c*) the kinds of attributes used to predict zinc content?

4. For each of the examples of GIS analysis given in Chapters 7 and 8, work out the sources of errors and develop flow charts to show how uncertainties propagate through the models.

Suggestions for further reading

Burrough, P. A. (1992). Development of intelligent geographical information systems. *International Journal of Geographical Information Systems*, 6: 1–15.

De Roo, A. J. P., Hazelhoff, L., and Heuvelink, G. B. M. (1992). Estimating the effects of spatial variability of infiltration on the output of a distributed runoff and soil erosion model using Monte Carlo methods. *Hydrological Processes*, 6: 127–43.

Fisher, P. F. (1991). Modelling soil map-unit inclusions by Monte Carlo simulation. *International Journal of Geographical Information Systems*, 5: 193–208.

Goodchild, M., and Gopal, S. (1989). *The Accuracy of Spatial Databases.* Taylor & Francis, London.

Heuvelink, G. B. M., Burrough, P. A., and Stein, A. (1989). Propagation of errors in spatial modelling with GIS. *International Journal of Geographical Information Systems*, 3: 303–22.

Taylor, J. R. (1982). *An Introduction to Error Analysis.* University Science Books, Oxford University Press, Oxford.

Thapa, K., and Bossler, J. (1992). Accuracy of spatial data used in Geographic Information Systems. *Photogrammetric Engineering and Remote Sensing*, 58: 841–58.

ELEVEN

Fuzzy Sets and Fuzzy Geographical Objects

Up to this point we have assumed that real world phenomena can be modelled either by exactly defined and delineated entities like polygons or by smooth continuous fields. Uncertainty has been treated probabilistically, using conventional statistical methods and spatial statistics. In this chapter we present ways of dealing with uncertainty, complexity, and vagueness in terms of fuzzy, overlapping sets. Instead of probability, fuzzy set theory uses concepts of admitted possibility, which is described in terms of the fuzzy membership function. Fuzzy membership functions permit individuals to be partial members of different, overlapping sets. The sets can be defined exogenously, using the Semantic Import model, or can be computed from multivariate data using the methods of fuzzy *k*-means. If the original data are measured at point locations, membership functions can be mapped by interpolation: zones where different fuzzy set surfaces intersect are locations of 'confusion' and can be extracted to provide crisp boundaries. The methods are illustrated with applications from soil survey, land classification, pollution mapping, and vegetation science.

Chapter 2 explained that in order to model geographical phenomena it is first necessary to divide the world either into crisp entities (land parcels, administrative units, soil mapping units, or ecotopes) or into continuous fields (atmospheric pressure, temperature gradients, groundwater levels, population density) that are discretized, usually as a regular grid. These fundamental spatial entities—points, lines, polygons, pixels—are described by their location, attributes and topology. A set of axioms (Chapter 2, pp. 28–9) provides the ground rules for manipulating spatial entities with the usual rules of logic, which includes mathematics. Paramount for any form of data retrieval and manipulation is the division of data into groups, or sets: those entities that match selection criteria, and those that do not.

In the process of classification and retrieval we unconsciously use the basic laws of thought, first developed by Aristotle, namely:

- The law of *identity* (everything is what it is—a house is a house)
- The law of *non-contradiction* (something and its negation cannot both be true—a house cannot be both house and not a house), and
- The *principle of the excluded middle* (every statement is true or false—this house is lived in or it is not).

The principle of the excluded middle ensures that all statements in conventional logic can have only two values—true or false, which can be coded as zero or one. This assumption lies at the heart of most of mathematics and computer science (Barrow 1992) so naturally the paradigm is very deeply embedded in our tools for computation and data retrieval. The principles of two-valued logic make class overlap, the concept of partial membership of a set, and the concept of partial truths impossible.

Many geographical phenomena, however, are not simple clear-cut 'entities'. The patterns produced by natural processes vary over many spatial and temporal scales and the ensemble entities are defined not by one but many interacting attributes. Consequently, it is often a very difficult practical problem to partition the real world into unique, non-overlapping sets. Figure 2.2 showed that when our perception of reality yields something that is neither clearly an entity nor a continuous field, we use simple pragmatism to force the observation into one or the other mode, because until recently we had no means in GIS, apart from statistics, for dealing with entities that are not crisply defined, nor for groups that are not mutually exclusive.

The realization that one could invent all manner of different, self-consistent forms of logical reasoning . . . had a liberating effect upon thinkers struggling with problems that seemed to defy traditional forms of argument. (Barrow 1992: 18)

Much effort has been expended on how best to define standards for geographical data, and the discussions lie at the heart of issues of interoperability (see Chapter 12). The fact is that by limiting the rules of logic to binary decisions we limit the retrieval and classification of data to those situations in which only a complete match is possible. In real life we often make compromises based on the *degree* with which an object meets our specifications—if the object is almost what we want we will gladly make do. For example, you want to buy a new house. Your specifications are that it must have three bedrooms, have a garden that is at least 300 m^2, is located within 10 minutes' drive of the station and shops, is no more than 20 minutes' walk from your work, and costs no more than £150 000. This sounds pretty exact, but what do you mean by '10 minutes' drive'? Is this an 'average' figure, and if so how is it measured? Can you specify the traffic conditions? Would you consider other options outside the specification, such as a house with 290 m^2 of ground costing £155 000 with four bedrooms if it matched other criteria?

Similar problems have not only bedevilled the definition of classification of natural geographical phenomena such as rock types, soil or vegetation classes, but also affect the derivation of socio-economic groupings, decisions in law courts, the definitions of nationality, and even the borders of the nation state (Burrough and Frank 1996).

Guilty, not guilty or not proven, m'lud? Some legal systems (e.g. in Scotland) provide for uncertainty, while others allow only binary decisions.

Even though we think we have defined classes exactly, we may not be able to assign individuals to the correct groups, either because the rules are ambiguous or measurements cannot be made with sufficient accuracy. Differences between soil surveyors on how to partition landscapes can lead to different data, and hence decisions based on them could be intrinsically unreliable (see e.g. Bie and Beckett 1973, Legros *et al.* 1996). The differences in interpreting the discretization rules adopted in data collection and classification persist through the database and have a direct effect on the presentation and interpretation of the results of modelling (Chapter 2).

Imprecision as a way of thought

Because formal thought processes in Western logic have traditionally emphasized the paradigm of truth versus falsehood, which is implemented in Binary or Boolean logic, we have very little formal training on how to deal with overlapping concepts. The law of the excluded middle and its role in mathematical proof—*reductio ad absurdum*—has been of paramount importance in scientific and philosophical development. The rules of logic used in computer query languages are all based on exact ideas of truth or falsehood, which implies that all objects in a database fit the same paradigm. In environmental data this is not necessarily so.

Conventional logic sometimes leads to unsolvable paradoxes. A statement such as 'Bongo says that he always lies' can be neither true nor false, because if the statement is true, it is incompatible with *always* lying. Such paradoxes can be solved by admitting partial truths, or by admitting class overlap, which has naturally happened in some cultures (e.g. see Barrow 1992, Burrough and Frank 1996). Also, we often allow class overlap and partial truths in natural language when we are not forced by scientific or legal considerations to work with exact concepts.

In natural language we have the full freedom to add modifiers to divide up the middle ground between potentially overlapping classes and any difficult situation can be decided locally. For example, consider the discrimination between tall and short people. At first, we might consider that a person who is 190 cm (6 ft 2.5 inches) in length is *tall*, and one who is 150 cm (4 ft 11 inches), *short* (Figure 11.1). But is a person who is 170 cm long, tall or short? If the tall/short class boundary is fixed halfway between these two extremes,

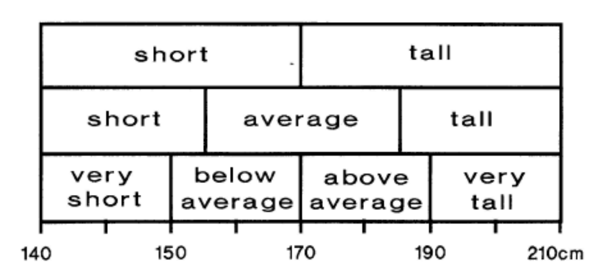


Figure 11.1. Natural language deals with fuzzy classes by redefinition and class splitting
In fuzzy sets, the grade of membership is expressed in terms of a scale that can vary *continuously* between 0 and 1

we cannot decide, so in practice we take the middle ground and create a new set of classes—*short—average—tall*. This creates new problems for classifying persons 154 or 186 cm in length, but we can easily create new classes in the areas of overlap to deal with the *below average* and *above average* cases. This process can continue as long as necessary until we have reached a generally acceptable system. Similar qualitative approaches allow us to deal with concepts such as baldness, the busyness of a holiday beach or the perception of climatic zones.

Many users of geographical information have a clear notion (or central concept) of what they need. Land evaluators and planners can usually define the ideal requirements for a particular kind of land use, but they are often unsure about just where the boundaries between 'suitable' and 'unsuitable' classes of land should be drawn. In many situations they also formulate their requirements generally in terms such as 'Where are areas of soil suitable for use as experimental fields', 'How much land can be used for smallholder maize?', 'Which areas are under threat of flooding?' or 'Which parts of the wetlands suffer from polluted discharges?' Such imprecisely formulated requests must then be translated in terms of the basic units of information available. Furthermore, not all information stored in the simple data model is exact, however. Many data collected during field survey are often described in seemingly vague terms: soils can be recorded as being 'poorly drained', having 'moderate nutrient availability', being 'slightly susceptible to soil erosion'; vegetation can be described as vital, partially vital, and so on. Even though standard manuals define these terms with more precision, in practice they retain a strong flavour of qualitative ambiguity.

In natural resource inventory and modelling it is not sensible to permit users to define their classes in a totally ad hoc way because we need to agree on a limited number of standardized classes to facilitate training, research, and general information transfer. Clearly, we need methods for specifying how we must deal with imprecise information and borderline cases.

Geographical phenomena and imprecision

Geographical phenomena are more complicated than many other multivariate defined entities because we

must consider grouping both in *attribute space* and in *geographical space*. Grouping in attribute space determines whether all entities are of the same kind, and the problem of determining a group identity is one of choosing a classification based on attributes that leads to unambiguous, non-overlapping classes with clear rules for allocation. Grouping in geographical space determines whether entities (or multivariate fields) of similar kind occupy contiguous regions.

Very often, when setting up taxonomies of natural or man-made entities, we have concentrated only on building class definitions from attributes, with the implicit, and not unreasonable, assumption that similar entities will cluster together. Very often this may be so, but there is a growing and large literature devoted to the study of short-range natural variation, which can be particularly acute in soils, groundwater quality, and other aspects of natural and cultural landscapes (Burrough 1993*b*).

In the following text we first examine how to deal with imprecision in overlapping attribute classes. We then illustrate how this understanding can be extended to geographical phenomena that are expressed either as continuous fields or as entities like polygons.

Fuzzy sets and fuzzy objects

Although other cultures have concepts of overlapping groups (see Barrow 1992), until the late nineteenth century Western thought had no formal means to handle such logical monsters. The wider introduction of ideas on multi-valued logic started in 1965 when Zadeh (1965) introduced the idea of 'fuzzy sets' to deal with inexact concepts in a definable way. The term 'fuzzy' has been seen by some as unfortunate, because it suggests an image of woolly, unstructured thought, which is therefore unscientific and bad. This is definitely not the case, but the name has stuck. Since the 1960s, however, the theory of fuzzy sets has been developed to the point where useful, practical tools are available for use in other disciplines (Zadeh 1965, Kandel 1986, Kauffman 1975).

Fuzziness is a type of imprecision characterizing classes that for various reasons cannot have or do not have sharply defined boundaries. These inexactly defined classes are called *fuzzy sets*. Fuzziness is often a concomitant of complexity. It is appropriate to use fuzzy sets whenever we have to deal with ambiguity, vagueness, and ambivalence in mathematical or conceptual models of empirical phenomena. If one accepts that spatial processes interact over a wide range of spatial scales in ways that cannot be completely predicted (cf. Chapters 6 and 10), then one can appreciate the need for the 'fuzzy' concept in geographical information. Fuzzy set theory is a generalization and not a replacement for the better-known abstract set theory which is often referred to as Boolean logic.

Fuzziness is *not* a probabilistic attribute, in which the degree of membership of a set is linked to a given statistically defined probability function. Rather, it is an admission of *possibility* that an individual is a member of a set, or that a given statement is true. The assessment of the possibility can be based on subjective, intuitive ('expert') knowledge or preferences, but it could also be related to clearly defined uncertainties that have a basis in probability theory. For example, uncertainty in class allocation could be linked to the possibility of measurement errors of a certain magnitude.

Crisp sets

Conventional or crisp sets allow only binary membership functions (i.e. TRUE or FALSE)—an individual is a member or it is not a member of any given set. Fuzzy sets, however, admit the possibility of PARTIAL MEMBERSHIP, so they are generalizations of crisp sets to situations where the class boundaries are not, or cannot be sharply defined, as in the case of tall and short people given above. The same is true for an environmental property such as 'internal soil drainage' which embraces all conditions from total impermeability to excessively free draining. To state just what is, and what is not, 'a moderately well-drained soil' requires not strict allocation to an exactly defined class, but a qualitative judgement that by implication allows the possibility of partial membership.

A *crisp set* is a set in which all members match the class concept and the class boundaries are sharp. The degree to which an individual observation z is a member of the set is expressed by the *Membership function*

MF^B, which for crisp (Boolean) sets can take the value 0 or 1. Note that z is used here as a general attribute value, as in regionalized variable theory (Chapter 6). Formally, we write:

$$MF^B(z) = 1 \quad \text{if } b_1 \leq z \leq b_2$$
$$MF^B(z) = 0 \quad \text{if } z < b_1 \text{ or } z > b_2 \qquad 11.1$$

where b_1 and b_2 define the exact boundaries of set A. For example, if the boundaries between 'unpolluted', 'moderately polluted', and polluted soil were to be set at $b_1 = 50$ units and $b_2 = 100$ units, then the membership function given in (11.1) defines all 'moderately polluted soils'.

Fuzzy sets

A *fuzzy set* is defined mathematically as follows: if Z denotes a space of objects, then the fuzzy set A in Z is the set of ordered pairs

$$A = (z, MF_A^F(z)) \quad \text{for all } z \in Z \qquad 11.2$$

where the membership function $MF_A^F(z)$ is known as the 'grade of membership of z in A' and $z \in Z$ means that z belongs to Z. Usually $MF_A^F(z)$ is a number in the range 0,1 with 1 representing full membership of the set (e.g. the 'representative profile' or 'type') and 0 non-membership. The grades of membership of z in A reflect a kind of ordering that is not based on probability but on admitted possibility. The value of $MF_A^F(z)$ of object z in A can be interpreted as the degree of compatibility of the predicate associated with set A and object z; in other words $MF_A^F(z)$ of z in A specifies the extent to which z can be regarded as belonging to A. So, the value of $MF_A^F(z)$ gives us a way of giving a graded answer to the question 'to what degree is observation z a member of class A?'. Figure

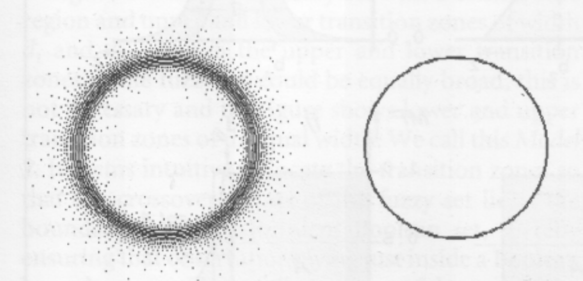


Figure 11.2. Images of fuzzy sets (left) and Boolean (crisp) sets (right)

11.2 uses the method of Venn diagrams to illustrate the difference between crisp and fuzzy sets.

Put simply, in fuzzy sets, the grade of membership is expressed in terms of a scale that can vary *continuously* between 0 and 1. Individuals close to the core concept have values of the membership function close to or equal to 1: those further away have smaller values. Note that this immediately gets around the problems of the principle of the excluded middle—truth is *not* absolute—and individuals can, to different degrees, be members of more than one set. The problem is to determine the membership function unambiguously.

Choosing classes based on attributes

In many cases the boundary values of crisp sets are chosen either (*a*) on the basis of expert knowledge (e.g. boundary values of discriminating criteria chosen by custom, law, or an external taxonomy), or (*b*) by using methods of numerical taxonomy. Classes based on expert knowledge are usually *imposed* or imported classes that are set up without direct reference to the local data set. They may approximate 'natural' divisions, but they are not optimal in any statistical sense. Only two parameters, the lower and upper boundary values, are needed. These classes are used a great deal in practical science and administration.

So-called natural classification methods produce classifications which are locally optimized to match the data set. In practice, the choice of classification method, parameter values, etc. can strongly affect the results of the classification. Methods of numerical taxonomy are mainly research tools that are data driven.

Both options are also possible with fuzzy sets. The first and simpler approach uses an a priori membership function with which individuals can be assigned a membership grade. This is known as the Semantic Import Approach or Model (SI). This is an analogue of (*a*) above.

The second is analogous to cluster analysis and numerical taxonomy in that the value of the membership function is a function of the classifier used. One frequently used version of this model is known as the method of fuzzy k-means.

Both methods can be usefully applied to environmental data, as is explained in the rest of this chapter. The semantic import model and other principles of fuzzy logic are described first, and then the method of fuzzy k-means.

Choosing the membership function: 1. The semantic import approach

The semantic import approach is useful in situations where users have a very good, qualitative idea of how to group data, but for various reasons have difficulties with the exactness associated with the standard Boolean model. The choice of fuzzy sets does not mean that one is opting out. The selection of boundaries for crisp sets and of class intervals can be an objective or a subjective process (e.g. see Burrough 1986, Evans 1977) depending on the way in which scientists agree to define classes: very often much thought goes into selecting sensible boundaries between class intervals. The same is just as true for assigning the membership function of a fuzzy set. The membership function should ensure that the grade of membership is 1.00 at the centre of the set, that it falls off in an appropriate way through the fuzzy boundaries to the region outside the set where it takes the value 0. The point where the grade of membership = 0.5 is called the '*crossover point*'. The membership function must be defined in such a way that these conditions hold, so not all functions are possible.

SUITABLE MEMBERSHIP FUNCTIONS FOR USE WITH THE SI APPROACH

Just as there are various type of *probability distribution* (normal, lognormal, rectangular, hyperbolic, Poisson, etc.), so there can be different kinds of fuzzy membership functions. These are used to determine the membership value at the edges of the set. Most common are the linear MF^F and the 'sinusoidal' MF^F.

The *linear* MF^F is given by a pair of sloping lines that peak at MF = 1 for the central concept of the set, *c*, and have *MF* values = 0.5 at the boundaries (Figure 11.3*b*). The slope of the line gives the width of the fuzzy transition zone. Note that areas inside the sloping lines, but outside the Boolean rectangle are zones of *partial truth.*

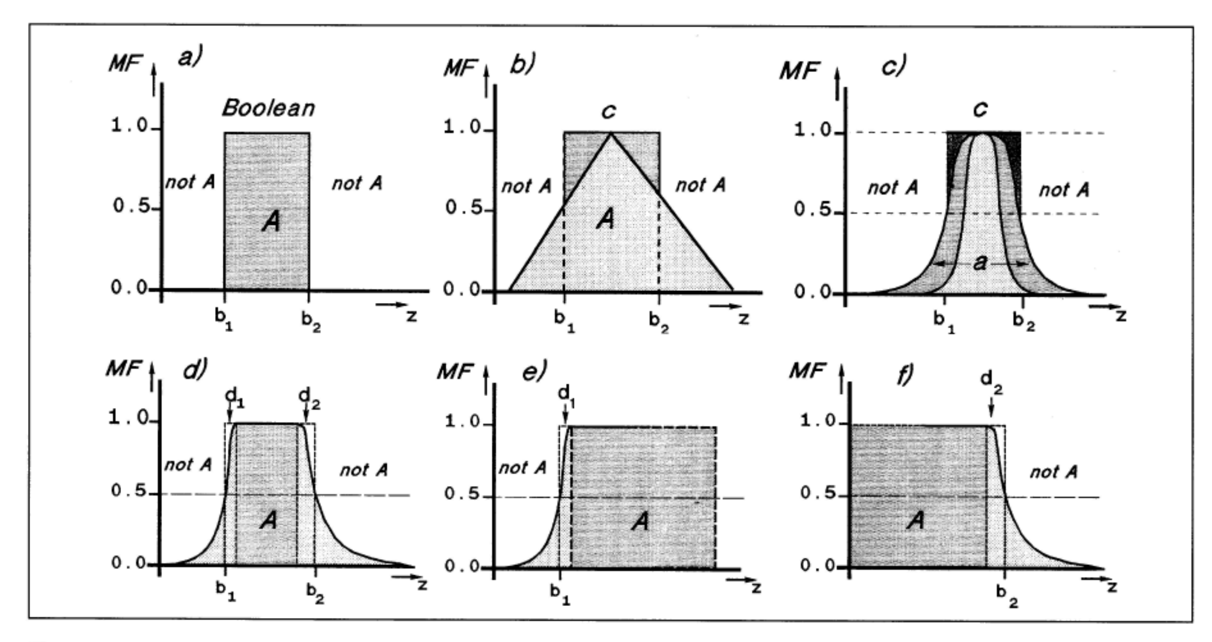


Figure 11.3. Boolean and fuzzy membership functions for the SI method

The *Sinusoidal* MF^F is given by:

$$MF_A^F(z) = \frac{1}{\left(1 + a(z - c)^2\right)} \quad \text{for } 0 \leq z \leq P \qquad 11.3$$

where A is the set in question, a is a parameter governing the shape of the function and c defines the value of the property z at the central concept. By varying the value of a, the form of the membership function and the position of the crossover point can be fitted to Boolean sets of any width (Figure 11.3*c*). We call this membership function 'Model 1'.

For example, consider the universe of soil depths from 0 to 200 cm, in which we wish to distinguish 'deep' soils from 'shallow' and from 'very deep' soils. If we chose $c = 100$ cm as the ideal centre (or 'standard index') of the fuzzy set of deep soils, then a value of $a = 0.0004$ gives lower and upper crossover points of 50 cm and 150 cm respectively where $MF_A^F(z) = 0.5$. The values of z flanking the central concept from the cross-over points (in this case from 50–100 cm and 100–50 cm) can be thought of as the *transition zones* surrounding the central concept of the fuzzy set.

Extending the Definition of Membership Functions to Sets with a Range of Values that Meet the Central Concept

Often, it may be sensible to extend the central concept of a fuzzy set to include a range of possible values rather than a single value, and this idea is implicit in the definition of crisp sets (equation 11.1—Figure 11.3*a*). Equation 11.3 can easily be modified to handle central concepts that cover a range of values; instead of using the parameter a to define the form of the fuzzy membership function it is easier in practice to use transition zones of width d_1 and d_2.

Figure 11.3*d* shows a fuzzy set with a central core region and upper and lower transition zones of width d_1 and d_2. Though the upper and lower transition zones to the fuzzy set could be equally broad, this is not necessary and the figure shows lower and upper transition zones of unequal width. We call this *Model 2*. It seems intuitive to locate the transition zones so that the crossover points of the fuzzy set lie at the boundaries of the equivalent Boolean set, thereby ensuring that observations lying just inside a Boolean boundary are still not full members of the core of the set. This is sensible if observations have an associated error band which means that though the recorded value may lie within the Boolean set, the true value may be outside. Therefore, values of d_1 and d_2 are essentially *half-widths* of the transition zones because they give the width *inside* the Boolean boundary to the value of z corresponding with the central concept. This also ensures that selections made by Boolean and fuzzy sets can be strictly compared and that at the boundaries, b_1, b_2, $MF^F = 0.5$.

Model 2 is defined by three equations:

$$MF^F(z) = \frac{1}{1 + \left(\dfrac{z - b_1 - d_1}{d_1}\right)^2} \quad \text{if } z < b_1 + d_1 \qquad 11.4a$$

$$MF^F(z) = 1 \quad \text{if } b_1 + d_1 \leq z \leq b_2 - d_2 \qquad 11.4b$$

$$MF^F(z) = \frac{1}{1 + \left(\dfrac{z - b_2 + d_2}{d_2}\right)^2} \quad \text{if } z > b_2 - d_2 \qquad 11.4c$$

where $MF^F(z)$ is the value of the continuous membership function corresponding to the attribute value z. Note that if parameters d_1 and d_2 are zero equation (11.4) yields the Boolean membership function (equation 11.1).

Asymmetric Membership Functions

In many situations only the lower or upper boundary of a class may have practical importance. This could be true, for example, in situations where we only wish to know if the soil is deep enough for a given purpose—if it exceeds this depth by 5 cm or 200 cm is immaterial. In these situations it is easy use the various parts of equation 11.4 to describe the membership function on the lower or the upper side (Model 3, Figure 11.3*e* or Model 4, Figure 11.3*f*).

Choosing Values for the Width of the Transition Zones

There seem to be no hard and fast rules for choosing the value of the d_1 and d_2, but it seems sensible to relate the width of the transition zones to what is known about the precision of measuring the attribute of the phenomenon or object in question. For data measured at points, the width of the transition zones could reflect the known accuracy of the measurement technique: for grid data interpolated by kriging, the width of the transition zone could be given by the kriging standard error. If fuzzy membership functions are used to represent how diffuse geographical boundaries may be (cf. Perkal's epsilon—see Chapter 9), the

widths of the transition zones of membership functions related to geographical boundaries could be defined using expert knowledge from the terrain.

FUZZY GEOGRAPHICAL OBJECTS

Before going further it is necessary to remind the reader that Fuzzy set theory initially only deals with diffuse boundaries and class overlap in attribute space. Applications to attributes of geographical entities is therefore straightforward. But in geographical information we must also deal with diffuse geographical boundaries, and fuzziness in geographical objects can also apply to features such as the boundaries of polygons, or to variation in membership function values that can be interpolated from point data. In this case, it is useful to refer to the continuously varying surface as a *fuzzy field*. As with all geographical data, the ease with which fuzzy information can be mapped depends on the strength of the spatial correlation structures, as explained in Chapter 6.

A PRACTICAL EXAMPLE

To demonstrate the differences between Boolean and fuzzy sets, Figure 11.4*a* shows a map of the clay content of the C horizon of the soil of part of the Lacombe Experimental Farm, Alberta interpolated by ordinary block kriging (Burrough *et al.* 1992). The area mapped measures 755 × 545 m with a 5 m resolution. Figure 11.4*b* shows the areas matching the Boolean class limits of ≥ 30 per cent clay and Figure 11.4*c* shows the fuzzy classification using Model 3 with a lower transition zone width of 5 per cent clay (equivalent to a kriging standard deviation of 5 per cent clay). Box 11.1 gives the pseudo code for computing these maps and shows how easy it is to compute the fuzzy membership function values.

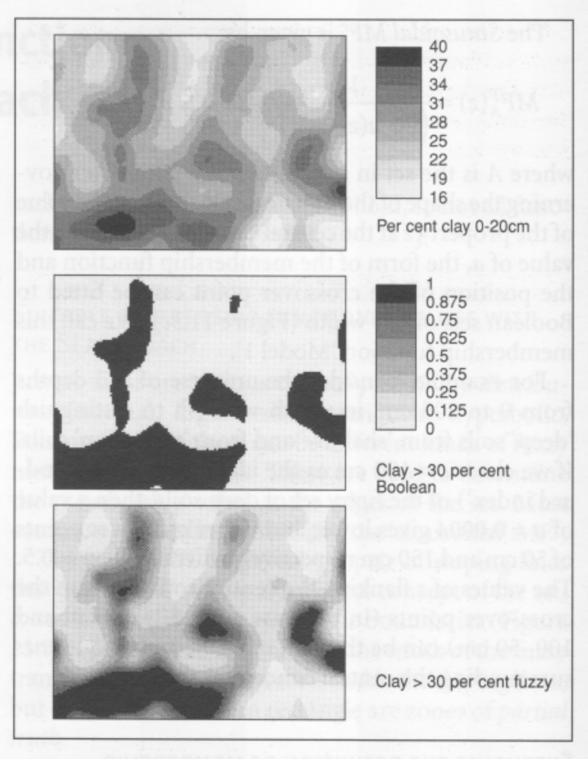


Figure 11.4. Boolean and fuzzy retrieval of soil texture

THE SEMANTIC IMPORT APPROACH: RECAPITULATION

Unlike the simple Boolean set where membership values are discretely 0 or 1, the Semantic Import approach transforms the data to a continuous membership function ranging from 0 to 1. The value of the membership function gives the *degree* to which the entity belongs to the set in question. Fuzzy classification is thus able to deal with the problem of unrealistic sharp class boundaries in a simple and intuitive way by defining the class limits and their dispersion indices.

BOX 11.1. PROCEDURES FOR COMPUTING BOOLEAN AND FUZZY MAPS

Computing maps of soil with clay content ≥ 30 per cent.

Boolean:

```
CLAY_30.BOO = IF(CLAY.EST GE 30 THEN 1 ELSE 0)
```

Fuzzy:

```
CLAY_30.FUZ = IF(CLAY.EST GE 35 THEN 1 ELSE (1/(1+((CLAY.EST−30−5)/5)**2)))
```

where CLAY.EST is the interpolated map of soil clay content.

Compared with Boolean classification, where only the boundary values have to be chosen, the extra problems the semantic import approach brings are (*a*) choosing between a linear, sinusoidal, or other function for assessing class membership and (*b*) selecting the values of the dispersion indices d_1 and d_2. Therefore the Semantic Import approach needs more information than the conventional Boolean method, but this is amply repaid by the extra sensitivity in data analysis, as will be demonstrated below.

Operations on several fuzzy sets

Just as with Boolean sets, data in fuzzy sets can be manipulated using logical query methods to select and combine data from several sets, and standard query languages in relational database management systems have been modified to accept continuous logical operations (e.g. Kollias and Voliotis 1991). The basic operations on fuzzy subsets are similar to and are a generalization of the AND/OR/NOT/XOR and other operations used for Boolean sets. Readers are referred to Kandel (1986), Kauffman (1975), or Klir and Folger (1988) for more details.

Box 11.2 gives the main logical operations for fuzzy sets. Note that the integral sign does *not* mean 'summation' but 'for the universe of objects *Z*'. For our purposes, the most used operations will be union (maximize), intersection (minimize), negation (complement), and the convex combination (weighted sum). The concept of the convex combination is useful when linguistic modifiers such as 'essentially' and 'typically' are to be used or when different attributes can compensate for each other. This idea is particularly appropriate in land evaluation as we shall see. Union, intersection, or convex combination lead to the computation of a new *MF* value, which we call the *Joint membership function* value or *JMF*.

It is now easy to see how an individual entity can be a member of one or more sets. Figure 11.5 shows the membership functions for two adjacent classes of clay texture of soil. Observation 1 is clearly in class *A*, and has a membership value of 1.0 for class *A* and 0.01 for class *B*. Observation 2 is near the class boundaries: it has a membership value of 0.4 in class *A* and a membership value of 0.7 in class *B*. The *JMF* for union (OR) for observation 1 is 1.0, and for 2 is 0.7; for intersection (AND) the *JMF* is 0.01 for observation 1 and 0.4 for observation 2. Note that if the two classes have different transition widths, then the membership values for a given observation do not have to sum to 1 for the Semantic Import approach. The same is true for classes defined using different attributes.

JOINT MEMBERSHIP VALUES FOR INTERPOLATED ATTRIBUTES

Box 11.3 demonstrates how union and intersection work with classes of different attributes. Suppose we have data from soil profiles that give the clay content for three layers—0–20 cm, 30–40 cm and 70–80 cm to an accuracy of ±5 per cent. Let us define a 'heavy textured soil' as one with a clay content of ≥30 per cent in the first layer, ≥35 per cent in the second layer, and ≥40 per cent in the third layer. As before we use a sinusoidal function with a dispersion value of 5 per cent (Model 3). In practice it may be necessary to distinguish profiles in which only one of the three horizons is 'heavy' (clay anywhere in the profile) from profiles where all three layers are heavy (clay throughout). The first situation is a union (OR), and requires selection of the *maximum* MF value: the second is a union (AND) and requires selection of the *minimum*. The right-hand columns of the table show which profiles would be selected, and which rejected by both categories. Note that if the measurement error is indeed ±5 per cent, measured clay values of 5 per cent less than the Boolean boundary (corresponding to a $MF^F = 0.2$) would be possible members of the set of 'heavy soils', though they would be completely rejected by the Boolean selection.

Figure 11.6 shows the same analysis, but now applied to interpolated surfaces of the clay content for the Lacombe data to distinguish the situation of 'clay in some part of the area' (Figure 11.6*a*,*b*) from 'clay throughout the profile' (Figure 11.6*c*,*d*). Clearly, fuzzy classification is a trivial operation in geographical information systems that provide 'cartographic algebra'. Note that the operation would be essentially the same were the clay content of the three layers expressed as soil polygons, but the appearance of the result would be dictated by the form of the mapping units.

If data are originally on a point support, one must decide whether first to interpolate all data to a com-

BOX 11.2. OPERATIONS ON FUZZY SETS

1. Two fuzzy sets, A and B, are said to be *equal* ($A = B$) iff (where iff means if and only if)

$$\int_Z MF_A(z)/z = \int_Z MF_B(z)/z \quad MF_A = MF_B$$

2. A is *contained* in B (A B) iff

$$\int_Z MF_A(z)/z \leq \int_Z MF_B(z)/z$$

3. The *union* of fuzzy sets A and B is the smallest fuzzy subset containing both A and B

$$A \cup B = \int_Z (MF_A(z) \vee MF_B(z))/z$$

v is the symbol for max. Union corresponds to the connective 'OR'.

4. The *intersection* of fuzzy sets A and B is denoted by $A \cap B$ and is defined by:

$$A \cap B = \int_Z (MF_A(z) \wedge MF_B(z))/z$$

^ is the symbol for min. Intersection corresponds to the connective 'AND'.

5. The *complement* of A corresponding to NOT is denoted by A^{-1} and is defined by:

$$A^{-1} = \int_Z (1 - MF_A(z))/z$$

6. The *product* of A and B (a 'soft' AND) is defined by:

$$AB = \int_Z (MF_A(z) \cdot MF_B(z))/z$$

7. The *bounded sum* of A and B (a 'soft' OR) is defined by:

$$A \oplus B = \int_Z^1 \wedge\ (MF_A(z) + MF_B(z))/z$$

where + is the arithmetic sum.

8. The *bounded difference* is defined by:

$$A \ominus B = \int_Z^0 \vee\ (MF_A(z) - MF_B(z))/z$$

9. If $A_1, \ldots A_k$ are fuzzy subsets of Z, and $w_1, \ldots w_k$ are non-negative weights summing to unity, then the *convex combination* of $A_1, \ldots A_k$ is a fuzzy set A whose membership function is the weighted sum:

$$MF_A = w_1 MF_{A1} + \ldots w_k MF_{Ak}$$

$$= \sum_{j=1}^{k} w_j MF_{Aj} \quad \text{where} \sum_{j=1}^{k} w_j = 1, \quad w_j > 0.$$

mon grid before carrying out fuzzy classification, or to classify the profile data first and then interpolate the membership functions. The second option is difficult for Boolean classes unless they are interpolated as indicator functions (see Chapter 6), but not for fuzzy *MF* values, though there may be theoretical problems (see Gruijter *et al.* 1997) because the fuzzy *MF* is not normally distributed or is constrained to the range 0–1. As in other areas of modelling, computational factors may decide which route is the best. For many users of geographical information systems it will be easier to apply the classification models to data that have already been interpolated.

Fuzzy classification with ordinal data

In Chapter 7 we showed how conventional Boolean logic is used in the 'top-down' classification of areas

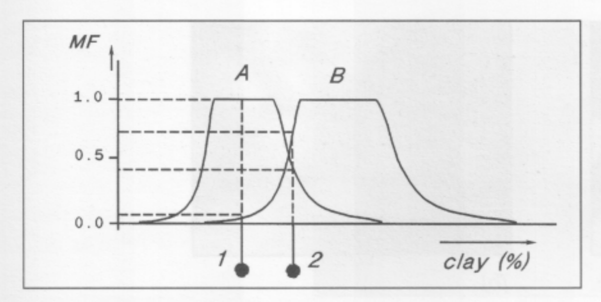


Figure 11.5. Two overlapping fuzzy membership functions

of land in terms of suitability to grow a crop. In this case the crop was maize and the location a small part of Kisii District in Kenya (Figure 7.10). Readers will recall that once the *land characteristics* (soil depth, soil type, slope classes) had been mapped, *land qualities* (water availability, oxygen availability, nutrient availability, and erosion hazard) were derived on a simple ternary ranking of 'poor' (or insufficient), 'moderate' (or just about sufficient), and 'good' (easily sufficient). The final land suitability classification was achieved by the simple rule of the most limiting factor. In other words, for every site or grid cell the worst ranking of any of the four land qualities determines the final ranking. The final result was that 84 per cent of the area was declared 'poor', 5 per cent 'moderate', and 11 per cent 'good'.

Inspection of the four maps of land qualities (Figure 7.10) shows that water availability and nutrient availability are the two land qualities most limiting for suitability for maize, yet these are the two attributes that can most easily be modified (by irrigation or by the addition of fertilizers or organic manures). As in the previous example with the clay content of the Lacombe soil, we can define a fuzzy membership function for each land quality. In this case the functions have been defined so that the transitions zones are equal to half a ranking. This distinguishes land qualities that are 'poor to moderate' from those that are 'moderate to good', for example.

Applying this idea to the Kisii land evaluation example yields the result presented in Figure 11.7. This reveals that far from 84 per cent of the area being 'poor', actually some 71 per cent scores better than being 0.7 of the ideal concept of 'good sites'. The implications for land use are obvious—with improved husbandry the landscape can be protected and be productive, not just written off.

Using fuzzy logic with continuous data such as drainage nets

As shown in Chapter 8 it is easy to derive secondary data from gridded continuous surfaces, slope, potential flow paths and upstream contributing areas. As in Chapter 8, a *wetness map* can be computed from a map of upslope contributing areas and a map of slopes at each grid cell.

$$W = \ln (UPL * g / \tan\beta) \qquad 11.5$$

where W is the estimated wetness of the grid cell, UPL is the number of upslope elements discharging

BOX 11.3. EXAMPLES OF MINIMA AND MAXIMA IN *MF* VALUES FOR CLAY

Site	Original data			Fuzzy memberships			Boolean MF		Fuzzy	
#	C1	C2	C3	C1	C2	C3	AND	OR	AND	OR
1	26.7	26.9	32.8	**0.27**	0.13	0.14	0	0	0.14	0.07
2	30.8	16.8	45.3	0.59	**0.04**	1.00	0	1	0.04	1.00
3	39.8	42.6	45.6	**1.00**	1.00	1.00	1	1	1.00	1.00
4	32.6	46.8	46.9	**0.81**	1.00	1.00	1	1	0.81	1.00
5	16.9	46.7	52.2	**0.07**	1.00	1.00	0	1	0.07	1.00
6	48.8	34.9	54.8	1.00	**0.49**	1.00	1	1	0.49	1.00
7	20.0	24.5	30.7	**0.10**	0.09	0.11	0	0	0.09	0.11

Comparative minima are in bold; comparative maxima underlined.
Limits: layer 1 ≥ 31%, layer 2 ≥ 35%, layer 3 ≥ 40%; widths 5%.

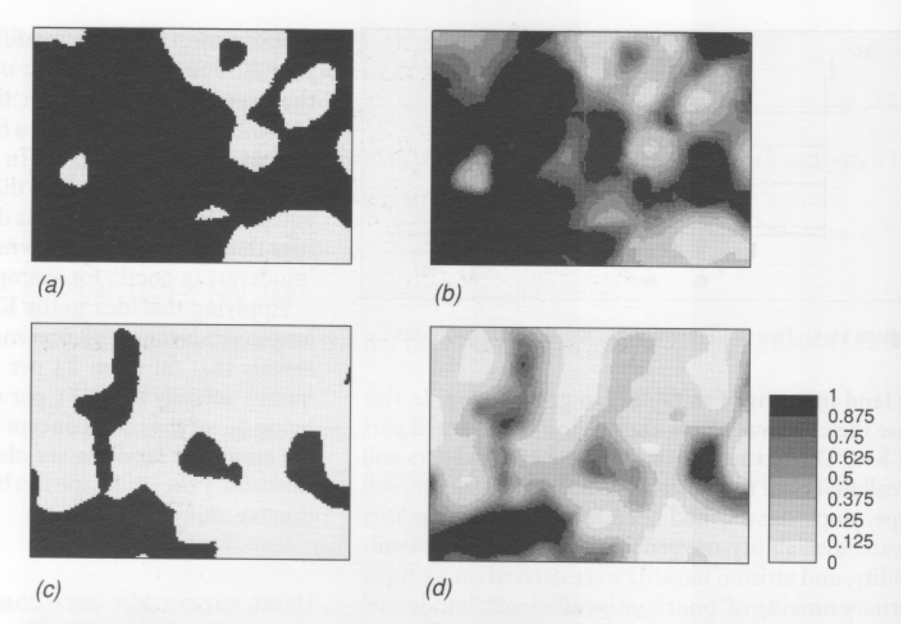


Figure 11.6. (*a,b*) Boolean and fuzzy union of sets to show where clay occurs anywhere in any layer; (*c,d*) Boolean and fuzzy intersection to show where clay occurs throughout the soil profile

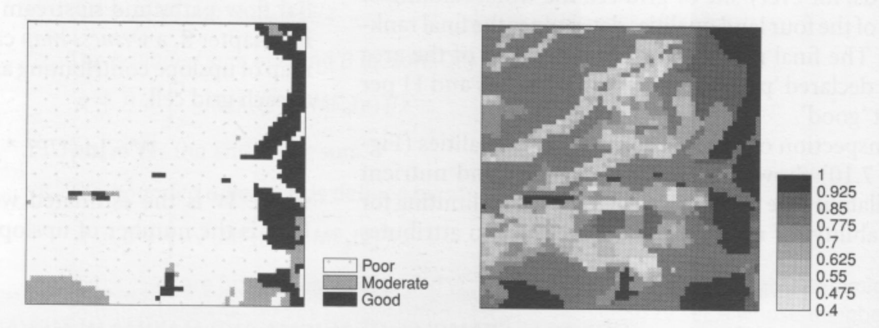


Figure 11.7. Comparison of results of top-down land capability classification using conventional rules (left) and the fuzzy equivalent (right)

through the grid cell, g is the area covered by each grid cell, and β is the slope of the land at the grid cell.

Because maps of upstream contributing areas and the wetness indices derived from them vary continuously there are no simple ways to extract features that could be termed 'boundaries'. Even contour lines do not make sense when the attribute in question is both continuous and clustered along limited pathways. Fuzzy sets are particularly appropriate for handling this kind of data, as the following example shows.

Figure 11.8 shows the derivation of a map showing where the higher values of the topologically derived wetness index combine with subsoil clay content. The limits of the clay content are taken as 40 per cent with a threshold of 5 per cent and the wetness index limit is 14 with a threshold of 6. Figure 11.8*c* shows the Boolean retrieval of all wetness zones with index >= 14 lying on subsoil clay >= 40 per cent and Figure 11.8*d* shows the fuzzy equivalent. Note that the fuzzy retrieval retains the whole of the drainage net in the areas of heavy clay, thereby preserving the topological continuity of the drainage, which is lost in the Boolean solution.

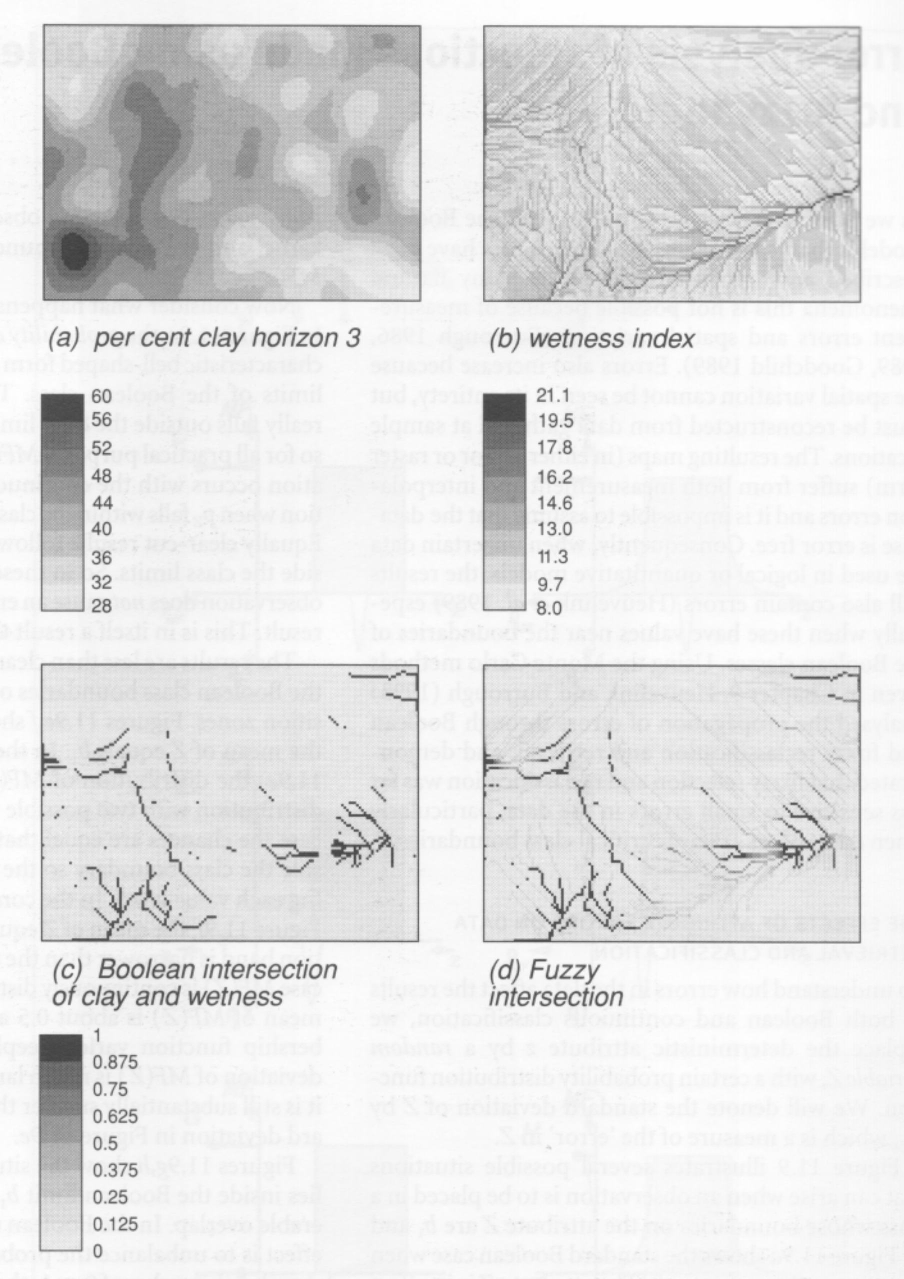


Figure 11.8. Using fuzzy methods for intersecting data layers with continuous data preserves detail and connectivity which is lost with Boolean selection

Error analysis of selections made using Boolean and fuzzy logic

As we have seen, a major problem with the Boolean model is that it assumes that the attributes have been described and measured exactly. In many natural phenomena this is not possible because of measurement errors and spatial variation (Burrough 1986, 1989, Goodchild 1989). Errors also increase because the spatial variation cannot be seen in its entirety, but must be reconstructed from data gathered at sample locations. The resulting maps (in either vector or raster form) suffer from both measurement and interpolation errors and it is impossible to assume that the database is error free. Consequently, when uncertain data are used in logical or quantitative models, the results will also contain errors (Heuvelink *et al.* 1989) especially when these have values near the boundaries of the Boolean classes. Using the Monte Carlo methods given in Chapter 9, Heuvelink and Burrough (1993) analysed the propagation of errors through Boolean and fuzzy reclassification and retrieval, and demonstrated that fuzzy selection and reclassification was far less sensitive to small errors in the data, particularly when data values are near critical class boundaries.

THE EFFECTS OF ATTRIBUTE ERRORS ON DATA RETRIEVAL AND CLASSIFICATION

To understand how errors in the data affect the results of both Boolean and continuous classification, we replace the deterministic attribute z by a *random variable* Z, with a certain probability distribution function. We will denote the standard deviation of Z by σ_Z, which is a measure of the 'error' in Z.

Figure 11.9 illustrates several possible situations that can arise when an observation is to be placed in a class whose boundaries on the attribute Z are b_1 and b_2. Figure 11.9*a* shows the standard Boolean case when there is no error, so $\sigma_Z = 0$. The attribute Z is in effect deterministic so the individual observation either falls entirely within the class boundaries (right-hand bar) or it falls outside (left-hand bar). The corresponding values of the membership function are 1 or 0 respectively. Because σ_Z is zero the membership value $MF(Z)$ is also error-free. Figure 11.9*b* shows the same situation where the individual observation is now classified by a continuous membership function. The right-hand observation is within the class kernel so $MF(Z) = 1$. The left-hand observation is outside the kernel and the Boolean boundaries so it returns an $MF(Z) < 0.5$.

Now consider what happens when σ_Z is non-zero. In Figure 11.9*c*, the *probability density* p_Z of Z takes the characteristic bell-shaped form and lies well within the limits of the Boolean class. The probability that Z really falls outside the class limits is vanishingly small so for all practical purposes $MF(Z) = 1$. The same situation occurs with the continuous membership function when p_Z falls within the class kernel (Figure 11.9*d*). Equally clear-cut results follow when p_Z lies well outside the class limits. So in these cases the error in the observation does *not* cause an error in the classification result. This is in itself a result that is worth noting.

The results are less than clear-cut when p_Z straddles the Boolean class boundaries or the continuous transition zones. Figures 11.9*e*,*f* show the situation when the mean of Z equals b_1. In the Boolean case (Figure 11.9*e*) the distribution of $MF(Z)$ becomes a discrete distribution with two possible values, 0 and 1. In this case the chances are equal that Z falls within or outside the class boundary so the probability of obtaining each value is 0.5. In the continuous case shown in Figure 11.9*f*, the mean of Z equals b_1 and the distribution band is narrower than the transition zone. In this case $MF(Z)$ is continuously distributed just as Z is. The mean of $MF(Z)$ is about 0.5 and because the membership function varies steeply at b_1, the standard deviation of $MF(Z)$ is much larger than for Z, though it is still substantially smaller than the Boolean standard deviation in Figure 11.9*e*.

Figures 11.9*g*,*h* show the situation when p_Z mainly lies inside the Boolean limit b_1, but still with considerable overlap. In the Boolean case (Figure 11.9*g*) the effect is to unbalance the probabilities of returning a membership value of 0 or 1; the probability of returning a value of 1 is increased. In the continuous case p_Z runs into the class kernel so the resulting distribution is a mixed distribution: it is the combination of a continuous and a discrete distribution. There is a definite chance that the membership value will be exactly 1 in some cases, though values less than 1 can also occur. Although not shown, the situation where p_Z is broader than the transition zone will generally yield mixed distributions of $MF(Z)$ with peaks at 0 and/or 1.

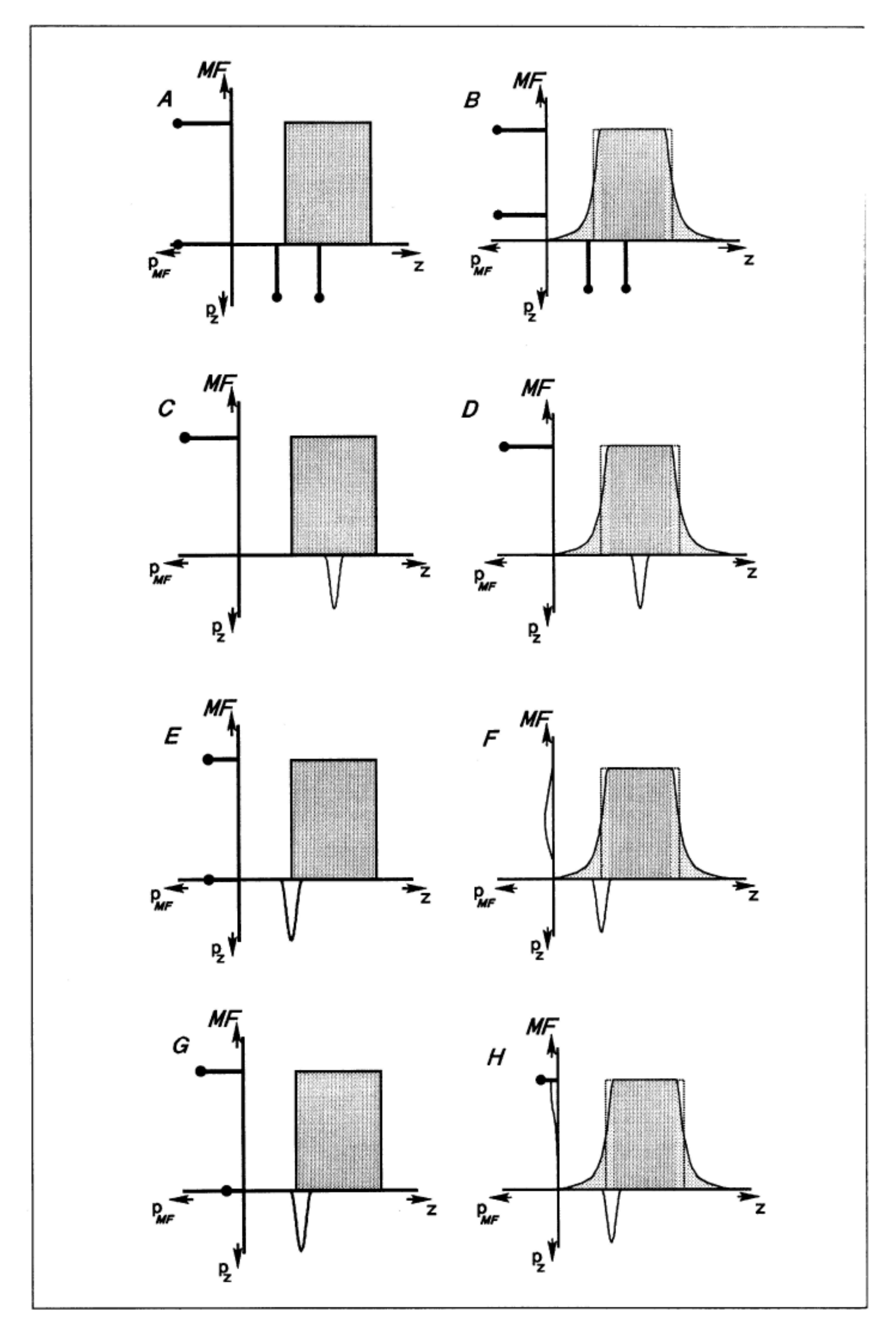


Figure 11.9. The effects of measurement errors on the results of Boolean classification (left column) and fuzzy classification (right column)

Applying the SI approach to polygon boundaries

So far we have concentrated on classifying attributes, but many kinds of mapping involve the delineation of similar areas of land, often by examining its external appearance, either in the field or from aerial photographs or satellite images. Observable features, such as colour changes, breaks of slope or edges of texture patterns, are used to infer 'boundaries'. The delineations are first drawn in pencil, then checked by making point observations within similarly classified areas, and then finalized in ink. Conventionally, the field sheets are redrawn as printing masters by cartographers who draw all soil, ecotope, geological boundaries with a line of standard width (usually 0.2 mm) irrespective of the original nature of the boundary in the field. These artifacts later digitized to yield digital choropleth maps on which, in theory at least, the drawn boundaries have zero width. This imposition of cartographic neatness suppresses useful information about the nature of spatial change and has misled generations of users of choropleth maps into thinking that soil, vegetation, or geological boundaries are always and everywhere sharp, and that the units are always homogeneous.

All field scientists know that spatial variations in soil, vegetation, or geology can occur abruptly or gradually. A dike intrusion can give rise to geological boundaries that are sharp at the scale of centimetres, but variations in texture of alluvial or aeolian deposits can occur over hundreds of metres or kilometres. It is not difficult to indicate whether a boundary identified during fieldwork or aerial photo interpretation is sharp or diffuse, because that can often readily be seen. It is also not difficult to attach an attribute indicating the relative or the estimated width of a boundary (sharp, medium, diffuse) to the arcs of a polygon when soil maps are digitized in a vector system, nor is it impossible to draw boundary lines in different styles to indicate their width—in fact all these have been possible for years (cf. Burrough 1980). The fact is that, apart from some concerns about the accuracy of boundary location, until recently few had thought it necessary and useful to include information *about the nature of the boundary* in a choropleth map (Lagacherie *et al.* 1996). Indeed, one objection to so doing was that it was impossible to compute the area of a polygon with diffuse boundaries.

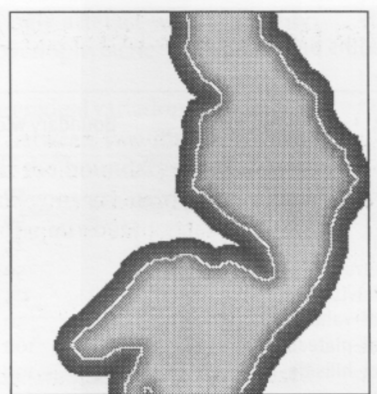
By using the methods of fuzzy logic on polygon boundaries it is very simple to incorporate information about the nature of those boundaries and also to calculate sensible area measures. There are at least two separate approaches, the map-unit approach and the individual boundary approach.

The map unit approach

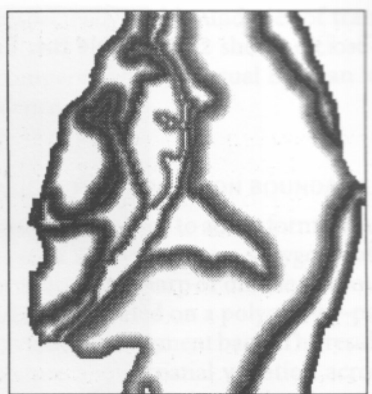
The simplest approach is to assume that a type of boundary (diffuse or abrupt) can be uniquely attributed to each kind of map unit or polygon. For example, soil units in narrow floodplains might have sharp, well-defined boundaries whereas boundaries between loess and coarser aeolian deposits might occur over several hundred metres. Information about the type of boundary can be converted to parameters for a fuzzy membership function of type Model 2, 3, or 4 (equation 11.4*a*,*b*,*c*) which is now applied to the distance from the drawn boundary. The widths of the transition zones must be chosen with the scale of the map in mind. For example, a sharp boundary on a 1 : 25 000 map drawn 0.2 mm thick covers 50 m so the width of the spatial transition zone centred over the drawn boundary location would be 25 m. A diffuse boundary at the same scale might extend over 500 m and have a transition zone width of 250 m.

Fuzzy transition zones can be computed from polygon boundaries by first spreading isotropically and outwards from the original delineation (cf. Chapter 7) and then applying a SI model to indicate the external gradation of MF value from well inside the polygon ($MF = 1$) to the outside. As an example, consider the soil map fragment in Figure 11.10 (taken from the Soil Map of the Mabura Hill Forest reserve in Guyana, Jetten 1994). Figure 11.10*a* shows the plateau phase of the plinthisols. This unit has fairly diffuse boundaries and as a first approximation let us assume that the true boundary has a locational uncertainty extending over 500 m, i.e. a transition zone of width 250 m.

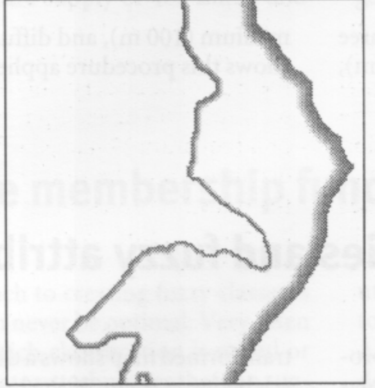
The procedure to compute a fuzzy boundary for a single polygon in a raster representation is simple. One extracts the boundary with an edge filter and then applies a spread command to compute the zones around the boundary (see Chapter 8). The parameters of the membership function are selected so that the locations corresponding to the original drawn boundary are at the crossover value, i.e. $MF = 0.5$. The membership function is then applied so that those sites well within



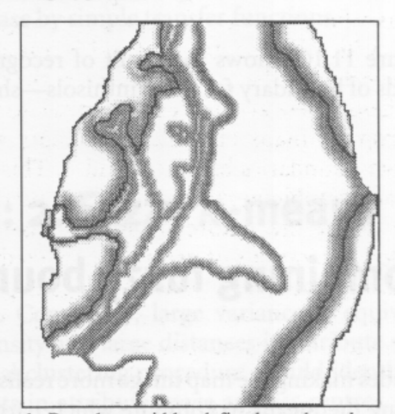
a. Plinthisols - plateau with diffuse fuzzy boundary (d_1 = 250m)



b. Boundary widths defined per polygon for all polygons.



c. Plinthisols - plateau, with local internal boundary widths of sharp (50m), medium (100m) and diffuse (250m).



d. Boundary widths defined per boundary section for units 1, 7 and 8

Figure 11.10. Using fuzzy membership functions to describe boundary widths

the original boundary receive a membership value of 1, those sites inside, but near the boundary receive a membership value between 0.5 and 1, and those sites outside the boundary receive a membership value below 0.5 concomitant with their distance from the boundary. Figure 11.10*a* shows the result, with the original boundary given in white. Clearly, it is now trivial to compute the area of the polygon that corresponds to any level of the membership function, so previously voiced doubts about estimating the area of polygons with fuzzy boundaries are no longer valid.

The procedure can be repeated for all map units, varying the width of the boundary to match the unit (Table 11.1). When all polygons are displayed with fuzzy boundaries it is not sensible to display both the inner and the outer zones of transition for them all because the resulting map is difficult to read. Therefore it is better to confine the display of the boundary *JMF* to the internal transition zone (*JMF* ge 0.5). Figure 11.10*b* shows the result for the different map units when the boundary widths have been assigned according to Table 11.1.

The individual boundary approach

Commonly, areas of soil or vegetation do not have the same kind of boundary all the way round the occurrence. For example, a raised river terrace may have a diffuse boundary on the upper side and a sharp boundary on the lower side where the river has cut it away. Other boundaries might be sharp for part of their length and diffuse for others. It is not difficult to add attributes to boundary sections and to spread these individually. The results are then pooled and fuzzy membership functions can be computed as before.

Table 11.1. Boundary widths per map unit

Soil unit	Boundary width d_1 (m)
Albic arenosol	50
Gleyic arenosol	50
Histosol	25
Haplic ferralsol	50
Dystric fluvisol	25
Plinthisol–valley	50
Plinthisol–plateau	100
Plinthisol–hillside	100

Figure 11.10*c* shows the result of recognizing three kinds of boundary for the plinthisols—sharp (50 m), medium (100 m), and diffuse (250 m): Figure 11.10*d* shows this procedure applied to more units.

Combining fuzzy boundaries and fuzzy attributes

Besides making the map image more realistic and providing the means to compute which parts of a polygon can really be regarded as part of the core area, fuzzy boundaries can also be useful when derived maps are made from original soil, geological, vegetation, or land use maps. Conventionally, derived maps are made by converting the map unit class into a number or code representing the representative value of the attribute in question. The result is that if adjacent map units have large, differing values of an attribute, the transformed map shows a discontinuity at the boundary between them. If the boundary is sharp, then the result is realistic, but if the boundary is really quite

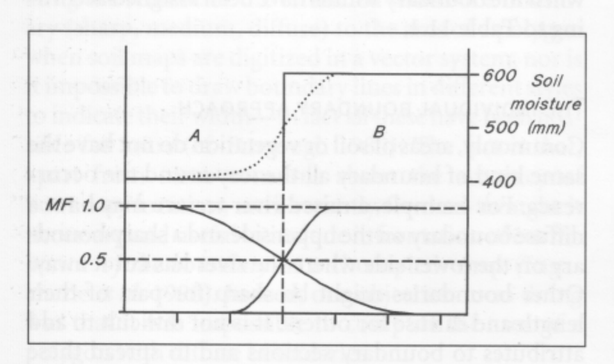


Figure 11.11. Using fuzzy overlap to compute weighted averages of attributes in the overlap zone

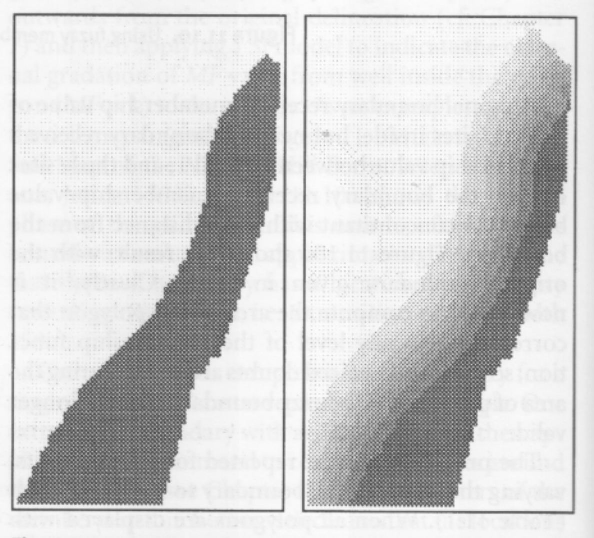


Figure 11.12. Conventional (left) and fuzzy (right) derived soil moisture maps for two adjacent polygons with a gradual, shared boundary

diffuse, the result is an ugly artefact which possibly can be removed by pycnophylactic interpolation (Chapter 5).

Information about the gradual variation of an attribute over the boundary between two dissimilar map units can be provided if the boundary membership functions of the two polygons are used to compute a weighted estimate of the attribute values over the boundary zone.

$$z_x = \frac{\Sigma(z_i{}^*MF_i)}{\Sigma MF_i} \qquad 11.7$$

Figure 11.11 shows the weighting procedure used to compute variation in soil moisture across the transition from Plinthisol plateau to hillside soil units, which have a soil moisture supply of 480 mm/a and 400 mm/a respectively and boundaries of transition width 250 m and Figure 11.12 shows a close-up of the results compared with the usual Boolean lookup table equivalence.

Recapitulation: fuzzy polygon boundaries

The SI approach can be used to add information about the abruptness of boundaries to a polygon database. Information about how sharp or diffuse the boundaries are can be incorporated on a polygon-by-polygon or on an individual line segment basis. The results give a sensible picture about spatial variation across and along boundaries which can be used to improve the information content of maps derived from the polygon database by simple transfer functions.

Choosing the membership function: 2. Fuzzy *k*-means

Although the SI approach to creating fuzzy classes is extremely flexible, it can never be optimal. Very often users may not know which classification is useful or appropriate, and exploratory techniques that can suggest suitable polythetic, overlapping classes can be useful. The method of fuzzy *k*-means (also known as fuzzy *c*-means) is such a technique which has been used in geohydrology (Frapporti *et al.* 1993), soil science (McBratney and de Gruijter 1993 and Odeh *et al.* 1990), and vegetation mapping (Jetten 1994). Several of these authors also prefer to call the method *continuous classification* rather than fuzzy classification, because of the continuity of the classes in attribute space, and also it is hoped, in geographical space.

The fuzzy *k*-means approach is primarily aimed at data reduction and convenient information transfer and was developed in the field of pattern recognition (Bezdek 1981, Bezdek *et al.* 1984). The method has strong relations with conventional methods of numerical taxonomy (Sneath and Sokal 1973). Data reduction is realized by translating a multiple attribute description of an object into *k* (or *c*) membership values to *k* classes or clusters. The clusters are optimal in the sense that the multi-variate within cluster variance is minimal. Near-zero variance means that all objects have nearly equal attributes, which means a high density and small distances between them in attribute space. Conversely, large variance is equivalent to low density and large distances in attribute space. An optimal clustering procedure should identify the dense spots in attribute space as class centres, while the boundaries between classes in attribute space should be located in the lowest density regions. Note that the fuzzy *k*-means approach initially says nothing about geographical contiguity of these optimal, overlapping classes.

Fuzzy *k*-means works by an iterative procedure that usually starts with an initial random allocation of the objects to be classified to *k* clusters (hence the name). Given the cluster-allocation, the centre of each cluster (in terms of attribute values) is calculated as the average of the attributes of the objects. In the next step, the objects are reallocated among the classes according to the relative similarity between objects and clusters. The similarity index is usually a well-known distance measure: the Euclidian, Diagonal (attributes are scaled to have equal variance), or Mahalanobis (both variance and co-variance are used for distance scaling) metrics are frequently used. Reallocation proceeds by iteration until a stable solution is reached where similar objects are grouped in one cluster (Figure 11.13).

Allocation of objects in conventional crisp *k*-means is always to the nearest cluster, with $MF = 1$ to this

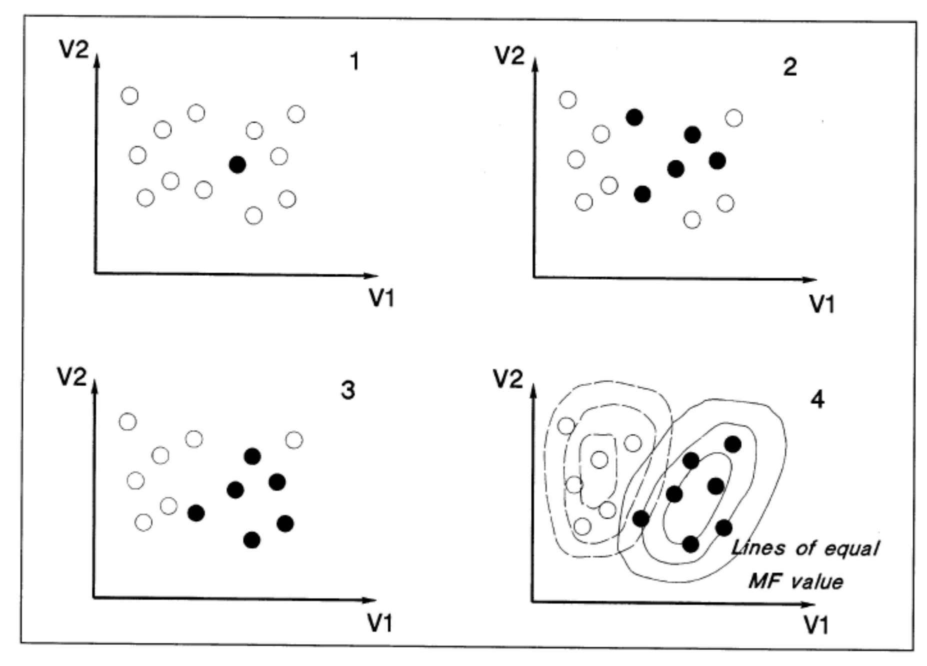


Figure 11.13. The formation of clusters using fuzzy *k*-means iteration

cluster and $MF = 0$ to all others. With fuzzy *k*-means, membership values may range between 0 and 1. Box 11.4 lists the algorithms used for calculating *MF* and the class centres. In ordinary *k*-means the *k* memberships of each object sum to 1. This causes a loss of degrees of freedom that may be compensated for by introducing an *extragrade class* (Gruijter and McBratney 1988). The parameter q is the so-called fuzzy exponent, which determines the amount of overlap in the cluster model. For $q \rightarrow 1$, allocation is crisp and no overlap is allowed. For large q there is complete overlap and all clusters are identical. Ideally q should be chosen to match the actual amount of overlap, which is generally unknown. Because distance is an overall measure of similarity, departure from the class centre of one attribute may be compensated by close correspondence of another. The effective relative weights of individual attributes in this valuation are determined by the type of distance measure used.

The net result of fuzzy *k*-means clustering is that individual multivariate objects (points, lines, or polygons) are assigned an MF value with respect to each of *k* overlapping classes. The centroid of each class is chosen optimally with respect to the data. In the variant of the technique used here, the MF values *sum to 1* rather than individually as is the case with the SI approach. This means that rather than all sets having equal value, as with SI, the sets are ranked according to importance. This has obvious implications for the procedures for manipulating sets given in Box 11.2 and means that procedures such as intersection and convex combination need to be carried out with care on the sets derived by fuzzy *k*-means.

Because the values of the MF_k are in effect new attributes, the geographical distribution of each set can be displayed by conventional methods of mapping, including interpolation. However, because of class overlap it is not always easy to read a map when more than interpolated surface is displayed, particularly when displays are limited to black and white. New methods for visualizing the results of fuzzy *k*-means classification on maps are clearly needed.

Examples of using fuzzy *k*-means and kriging to map overlapping, multivariate classes

(a) Heavy metal pollution on the Maas Floodplains Chapters 5 and 6 demonstrated interpolation methods with a data set of soil samples taken from a regularly flooded area of the River Meuse floodplain in the south of the Netherlands and data from a slightly larger area were used to illustrate costs and benefits in Chapter 10. Here we also use the larger subset of the data, but this time include all attributes (elevation,

BOX 11.4. ALGORITHMS USED IN FUZZY *k*-MEANS CLUSTERING

1. Membership μ of the *i*th object to the *c*th cluster in ordinary fuzzy *k*-means, with *d* the distance measure used for similarity, and the fuzzy exponent *q* determining the amount of fuzziness:

$$\mu_{ic} = \frac{[(d_{ic})^2]^{-1/(q-1)}}{\sum_{c'=1}^{k}[(d_{ic'})^2]^{-1/(q-1)}} \qquad 11.4.1$$

2. membership μ of the *i*th object to the *c*th cluster in fuzzy *k*-means with extra-grades, *d* and *q* as above, and with α as a weighing factor between extragrade and ordinary classes:

$$\mu_{ic} = \frac{[(d_{ic})^2]^{-1/(q-1)}}{\sum_{c'=1}^{k}[(d_{ic'})^2]^{-1/(q-1)} + \left[\frac{1-\alpha}{\alpha}\cdot\sum_{c'=1}^{k}[(d_{ic'})^2]^{-1/(q-1)}\right]} \qquad 11.4.2$$

3. membership μ of the *i*th object the extragrade cluster in fuzzy *k*-means with extra-grades:

$$\mu_{i*} = \frac{\left[\frac{1-\alpha}{\alpha}\cdot\sum_{c'=1}^{k}(d_{ic'})^2\right]^{-1/(q-1)}}{\sum_{c'=1}^{k}[(d_{ic'})^2]^{-1/(q-1)}] + \left[\frac{1-\alpha}{\alpha}\cdot\sum_{c'=1}^{k}(d_{ic'})^2\right]^{-1/(q-1)}} \qquad 11.4.3$$

4. *j*th attribute value *z* of the *c*th cluster in ordinary fuzzy *k*-means:

$$z_{cj} = \frac{\sum_{i=1}^{n}(\mu_{ic})^q * z_{ij}}{\sum_{i=1}^{n}(\mu_{ic})^q} \qquad 11.4.4$$

5. *j*th attribute value *z* of the *c*th cluster in fuzzy *k*-means with extragrades:

$$z_{cj} = \frac{\sum_{i=1}^{n}\left[(\mu_{ic})^q - \frac{1-\alpha}{\alpha}\cdot(d_{ic})^{-4}\cdot(\mu_{i*})^q\right] * z_{ij}}{\sum_{i=1}^{n}\left[(\mu_{ic})^q - \frac{1-\alpha}{\alpha}\cdot(d_{ic})^{-4}\cdot(\mu_{i*})^q\right]} \qquad 11.4.5$$

distance from the river, heavy metal concentrations (Cd, Zn, Pb, Hg), and organic material) to illustrate the multivariate fuzzy *k*-means approach.

Following the original flooding frequency map (see Chapter 5), the attribute data were classified into three classes with a fuzzy overlap $q = 1.5$. The fuzzy *k*-means classification yielded a new data set of memberships to each of the three classes for each site sampled. The scatter plot of membership values for the three classes (Figure 11.14) shows that the data are generally well separated in attribute space, with only a few observations scoring intermediate membership values. Table 11.2 presents summary statistics for the classes.

Figures 11.15*a*,*b*,*c* show the interpolated surfaces for the three classes. There is a strong suggestion of a relation between the three fuzzy classes and the flooding

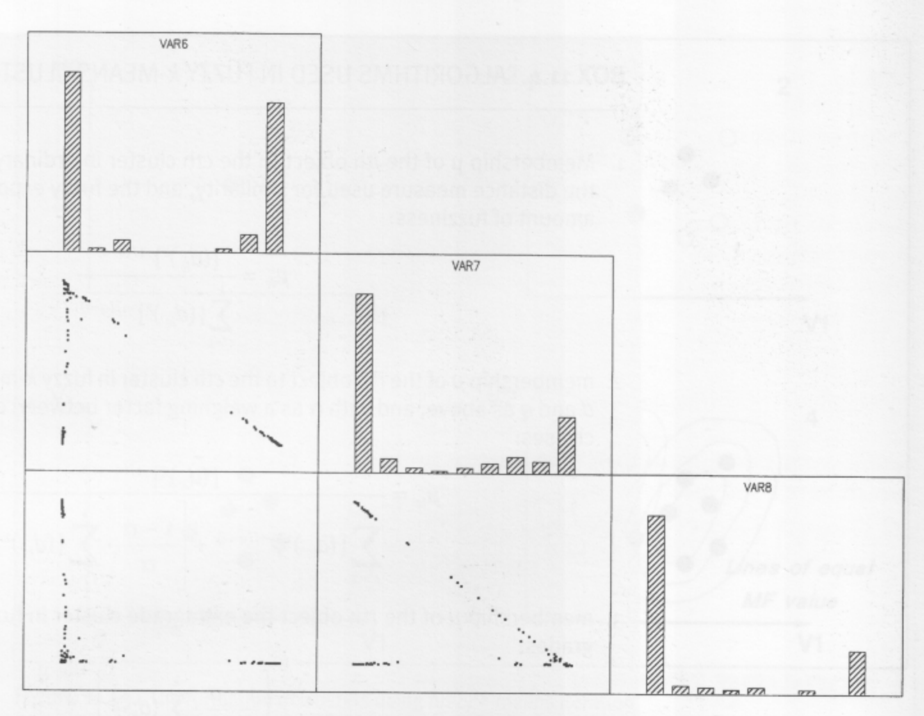


Figure 11.14. Scatter diagrams of membership values for soil pollution data

Table 11.2. Statistics of fuzzy classes, floodplain pollution

Class	Minimum membership	Maximum membership	Mean membership	Variogram type	Nugget C_0	Sill C_1	Range (m)	Ratio C_1/C_0
1	.0002	.999	.468	Sph	.0405	.267	1170	6.59
2	.001	.9983	.314	Sph	.0424	.0904	420	2.13
3	.0001	.9972	.217	Sph	.0590	.103	1170	1.75

frequency classes, namely that flood frequency class 1 covers similar areas as fuzzy class 3, flooding frequency class 3 covers a similar area to fuzzy class 1, and fuzzy class 3 occupies areas to the south and north. Note the distribution of these classes over the part of the area used in chapters 5 and 6, which is outlined in Figure 11.15*e*. In that area only two fuzzy classes are really important, which was one of the conclusions reached in the anovar analysis of the flood frequency mapping of the smaller area.

(b) Rainforest types A similar exercise was carried out on the results of a vegetation survey of 252 plots of each 0.05 ha on a 100 × 100 m grid in the rainforest in Guyana (Jetten 1994). The original species presence/absence data were first grouped into abundance classes before being analysed by a fuzzy *k*-means procedure which yielded three stable classes, termed 'Dry Evergreen Forest', 'Mixed Forest', and 'Wet Forest' (Jetten 1994). Table 11.3 presents summary statistics of these classes.

Class memberships were interpolated to a 50 m grid and the maps are shown in Figure 11.16. Unlike the pollution classes these are not spatially well correlated, as can be seen from the ratio of sill to nugget variances for the fitted variograms. Clearly, although the attrib-

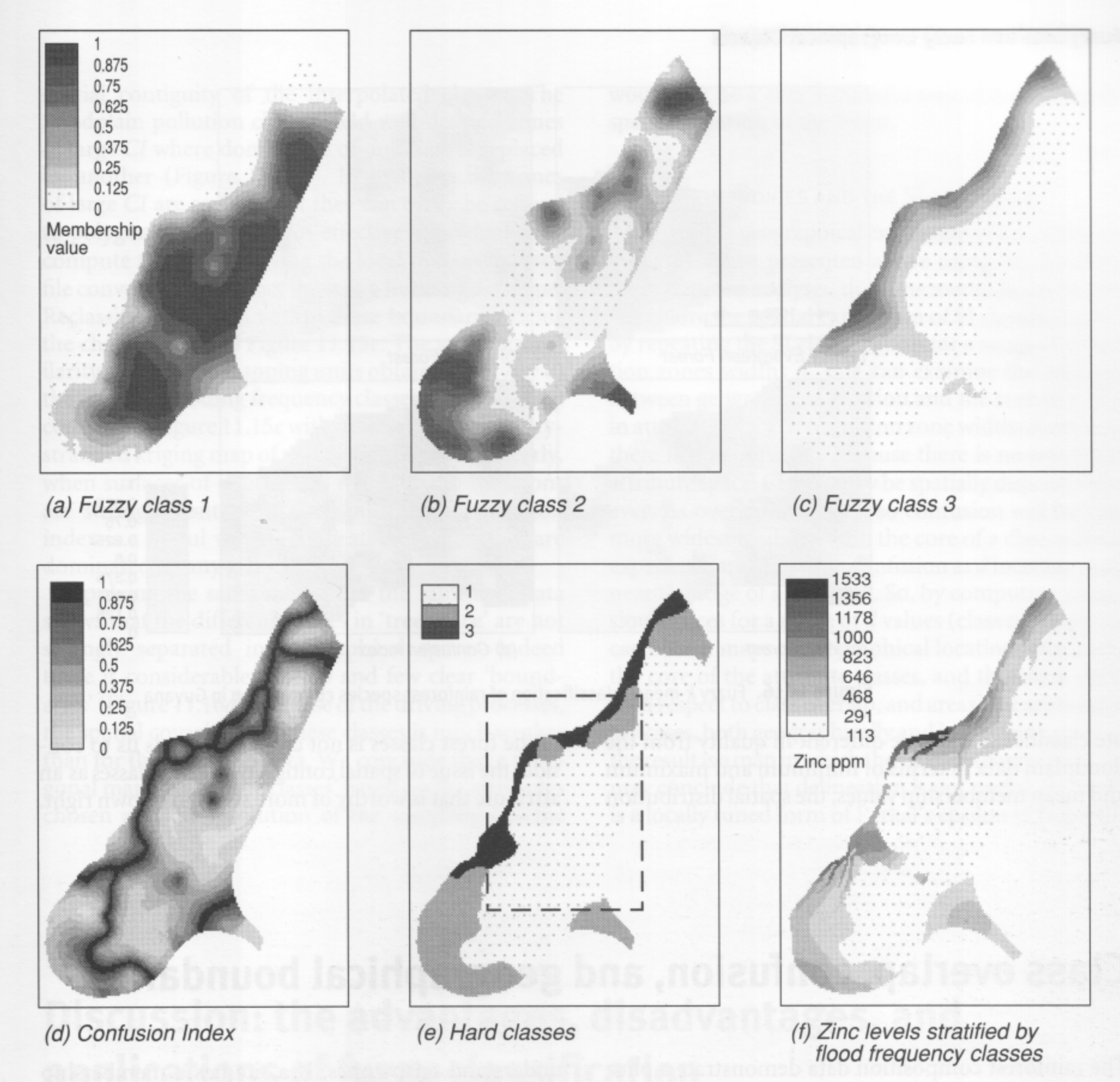


Figure 11.15. *(a,b,c)* Maps of interpolated fuzzy *k*-means membership values for the three soil pollution classes; *(d)* map of confusion indices; *(e)* crisp map showing where each class is dominant—the dashed line box outlines the area used in Chapters 5 and 6 to illustrate interpolation; *(f)* map of zinc levels obtained by stratified kriging for comparison with *(e)*

Table 11.3. Statistics of fuzzy classes, rainforest species

Class	Minimum membership	Maximum membership	Mean membership	Variogram type	Nugget C_0	Sill C_1	Range (m)	Ratio C_1/C_0
1	0.0	.9999	.288	Sph	.0250	.0400	216	1.60
2	0.0	1.0000	.445	Sph	.0500	.0518	1770	1.04
3	0.0	1.0000	.267	Sph	.0280	.0320	260	1.14

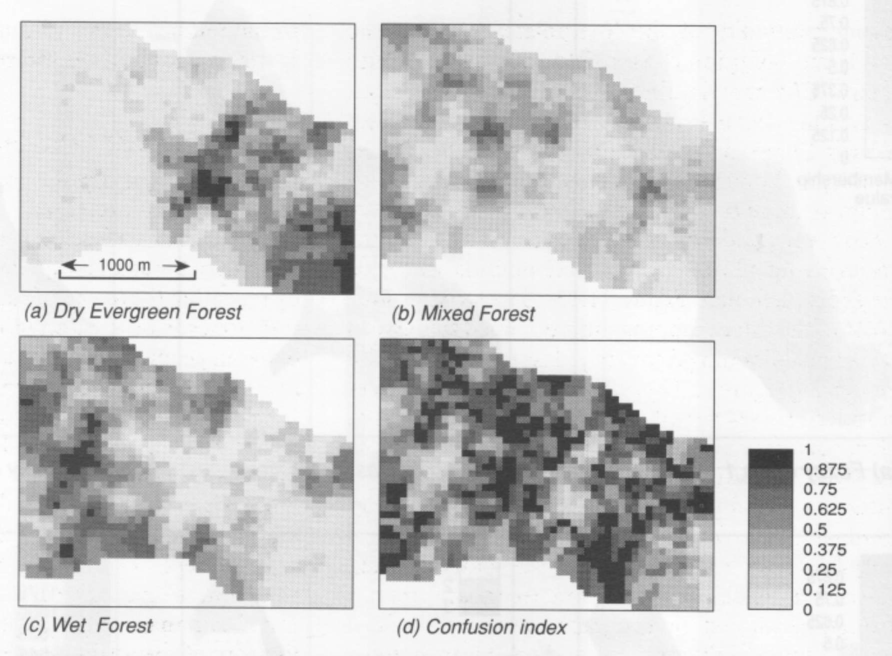


Figure 11.16. Fuzzy *k*-means classification of rainforest species composition in Guyana

ute classification is little different in quality from the floodplain data in terms of minimum and maximum and mean membership values, the spatial distribution of the forest classes is not clear. This leads us to consider the issue of spatial contiguity in fuzzy classes as an attribute that is worthy of more study in its own right.

Class overlap, confusion, and geographical boundaries

The rainforest composition data demonstrate a phenomenon that was mentioned at the beginning of this chapter, namely that data clustered in attribute space do not necessarily cluster in geographical space. From the parameters of the variograms of the fuzzy classes of both pollution data and rainforest data it is clear that the flooding process creates strong spatial correlation in the patterns of pollution, but the processes controlling the distribution of rainforest trees do not. This means simply, that mapping pollution zones alongside a flooding river is a lot easier than mapping rainforest stands. In the latter case, one tends to get easily confused about where the 'core areas' of each class of forest is situated—there seem to be lots of transition areas.

The concept of 'confusion' may help us to combine interpolated maps of fuzzy memberships into easy to understand crisp zones. If a site has a membership value near 1.0 in one of the fuzzy classes, it is clear to which class it belongs. But if the *MF* values for two or more classes are similar, it is by no means clear as to which class the site should be allocated. The situation is therefore confusing. Let us define the degree of class overlap in attribute space in terms of a *confusion index* for each observation, grid cell, or object; two possible forms are:

$$CI_1 = 1 - (MF_{max} - MF_{mx2}) \qquad 11.8$$

or

$$CI_2 = MF_{max}/MF_{mx2} \qquad 11.9$$

Using equation (11.8) to compute the *CI* for both sets of data provides interesting information on the

spatial contiguity of the interpolated classes. The floodplain pollution classes yield well-defined zones of large *CI* where dominance of one class is replaced by another (Figure 11.15*d*). In this case the zones of large *CI* are so thin that they can easily be refined to polygon boundaries. An effective procedure is to compute those cells having the local maximum profile convexity and extract them as a Boolean data type. Reclassifying all cells within these boundaries yields the choropleth map Figure 11.15*e*. The striking similarity of the three mapping units obtained this way to the original flooding frequency classes is provided by comparing Figure 11.15*e* with the flooding frequency-stratified kriging map of zinc (Figure 11.15*f*). Clearly, when surfaces of continuous membership functions are strongly contiguous, computing the confusion index is a useful way to delineate the zones that are dominated by any given class.

Applying the same method to the rainforest data shows that the different classes in 'tree space' are not strongly separated in geographical space—indeed there is considerable overlap and few clear 'boundaries' (Figure 11.16*d*). Because of the driving processes, the spatial covariance of these classes is much weaker than for the pollution data. We conclude that a polygonal mapping of 'tree classes' based on the attributes chosen and the resolution of the sampling scheme would not be a very successful way of describing the spatial variation of the forest.

Confusion indices and the SI approach

Although the geographical expression of the confusion index has been presented above using the results of fuzzy *k*-means analyses, the same methods can be used to explore the spatial expression of SI classes. Indeed, by repeating the SI classification for a range of transition zones widths d_i, one can examine the relations between geographical location and the core of a class in attribute space. If transition zone widths d_i are zero, there is no confusion, because there is no overlap in attribute space. Classes may be spatially disjoint, however. As overlap increases, so confusion will become more widespread. Areas in the core of a class will not experience an increase in confusion as d increase: areas near the edge of a class will. So, by computing confusion indices for a range of d values (class overlap) one can obtain maps of geographical locations that are in the core of the attribute classes, and therefore stable with respect to class overlap, and areas that are boundary cases, both geographically and in attribute space. If the result is a map in which the CI is low everywhere, we may conclude that delineation is not useful. So the CI is a locally tuned form of Perkal's epsilon (Chaper 9).

Discussion: the advantages, disadvantages, and applications of fuzzy classification

Advantages

The material presented in this chapter has demonstrated the advantages of both the SI approach and fuzzy *k*-means classification over conventional data retrieval and classification methods. It has been shown how exact classification loses information and increases the chance of classification errors when data are corrupted by inexactness, which is usually the case with environmental data. It has also shown that applying the SI approach to exactly delineated polygons can improve their information content, providing information about the nature of the sharpness or diffuseness of the identified boundaries is available.

There is a growing literature that illustrates the practical advantages of the SI approach for land classification and land evaluation (Altman 1994, Davidson *et al.* 1994, Wang *et al.* 1990), though in the scientific literature applications of fuzzy *k-means* classification have received more attention: for soil and environmental sciences see Gaans *et al.* (1986), Vriend *et al.* (1988), Odeh *et al.* (1990), Wang *et al.* (1990), Powell *et al.* (1991), McBratney and de Gruijter (1992), McBratney *et al.* (1992), Weiden *et al.* (1992), Gaans *et al.* (1992) and Frapporti *et al.* (1993), Burrough *et al.* (1997), Burrough and Frank (1996), Lagacherie *et al.* (1996), de Gruijter *et al.* (1997).

As with the SI approach, the overlapping classes and gradual boundaries resulting from the fuzzy *k*-means approach appear to be more congruent with reality than conventional methods. The fuzzy membership values carry through more of the initial information than crisp classification (McBratney *et al.* 1992). Membership values can easily be interpolated over space and the resulting variogram analysis and patterns demonstrate clearly if a given attribute class also has a coherent geographical distribution.

In contrast to the SI approach with predefined classes and class boundaries, the fuzzy *k*-means approach yields locally optimal classes which are not necessarily based on assumptions of linearity, unlike conventional data-reduction techniques, such as principal component and correspondence analysis. While linear correlations may sometimes be enhanced by means of suitable transformation functions (logarithmic, polynomial, etc.), attributes of soil, sediment, or water often co-vary in a complex, inherently non-linear manner. The fuzzy *k*-means approach captures this non-linear covariation (de Gruijter and McBratney 1988, Odeh *et al.* 1990). Moreover, both continuous and categorical attribute data can be easily combined (Odeh *et al.* 1990). In general, ordination techniques are sensitive to deviations from normality of the frequency distributions of the attributes, including bimodality and outlying values (Vriend *et al.* 1988). Because of the fuzziness allowed, the fuzzy *k*-means approach appears to deal accurately with both ends of the spectrum from co-variation to discrete clustering.

Disadvantages

As with most parametric methods, the greatest difficulties come with choosing the values of the control parameters to obtain the best results. With the SI approach, the user must choose the kind of membership function, boundary values, and transition widths. With fuzzy *k*-means using multiple attributes, the goodness of the fuzzy clustering obtained is difficult to visualize and evaluate. This relates to the choice of *k*: the number of classes, and *q*: the amount of overlap or fuzziness allowed. The added difficulty is that the optimal fuzziness may depend on the number of classes and vice versa. A formal approach using diagnostic functionals is possible (Bezdek 1981; Roubens 1982), but not always successful. Scientific insight or compliance with existing classification schemes may help (Vriend *et al.* 1988; McBratney and de Gruijter 1992; Frapporti *et al.* 1993). Other problems with fuzzy *k*-means are the choice of attributes and which distance measure to apply, which determine the degree and shape of fuzziness in the model and the way different attributes are compromised. It should be noted that the degree of fuzziness is always assumed equal for all classes.

While the fuzzy *k*-means procedure may seem to result in less arbitrary shaped fuzzy boundaries than used with the SI approach, this is only a matter because the method is less direct. Adequate cartographic techniques to map *k* classes (or *k* + 1extragrade) concurrently still have to be developed for both SI and fuzzy *k*-means (but see van der Wel and Hootsmans, 1993).

Conclusions—SI or fuzzy *k*-means?

The decision to use SI or fuzzy *k*-means approaches in continuous classification depends on the context of the problem and also on the level of prior information. In situations where a well-defined and functional

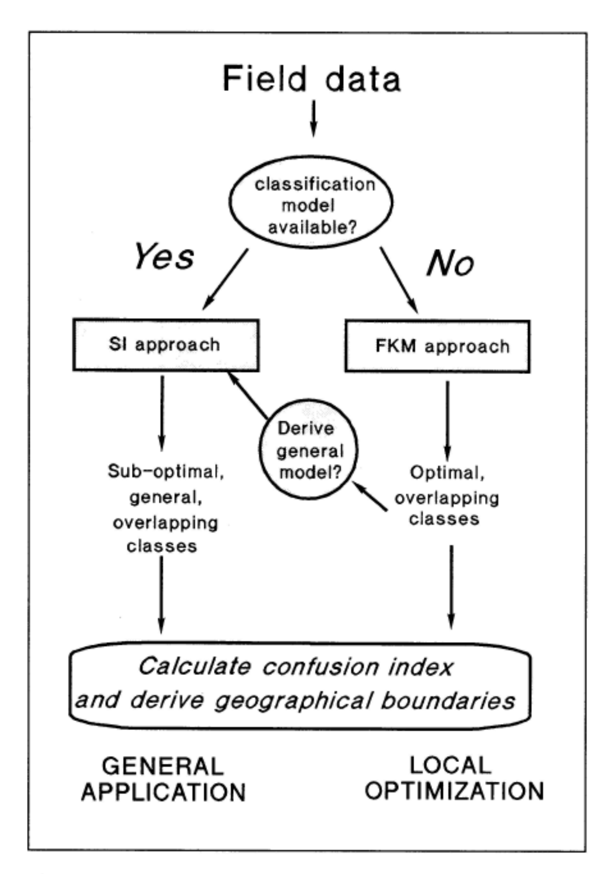


Figure 11.17. Flowchart governing the choice of fuzzy classification using either the SI or the fuzzy *k*-means approach

classification scheme exists, the straightforward SI approach offers great advantages but the precise definition of the fuzzy boundaries requires some special attention. Studies reported here have shown that SI continuous classes are more robust and less prone to errors and extremes than simple Boolean classes that use the same attribute boundaries. The results can be mapped showing the relations between gradual changes in attribute values/membership functions in both data space and geographical space.

In the fuzzy *k*-means approach, the criteria that distinguish between the ultimate classes are a result of the analysis rather than input to the model. These criteria may be complex non-linear functions of the original attributes, thus there is no clear-cut relation between membership values and attribute values. While objects having identical attribute values will end up with identical memberships, the reverse is not generally true because of the mutual compensation possible among attributes. Since data reduction is one of the main reasons for using class memberships instead of the original attribute values, some loss of information will always be faced.

The fuzzy *k*-means approach is appropriate when information about the number and definition of classes is lacking. Fuzzy *k*-means methods yield sets of optimal, overlapping classes that can also be mapped in data space and in geographical space. The results can be used as (*a*) a once-only analysis of a limited area to identify the lineage of (spatial) variability (cf. Vriend *et al.* 1988), (*b*) as a means to set up classes so that membership functions can be interpolated which in turn can predict attribute values at unsampled locations (McBratney *et al.* 1992), or (*c*) as a means better to define the simple membership functions that are to be used in straightforward SI analyses (Figure 11.17). To our knowledge no one has yet published an example of the latter, but the procedure parallels that using numerical classification and discriminant analysis by Burrough and Webster (1976) to set up and improve a reconnaissance classification for soil survey (Acres *et al.* 1975).

Questions

1. List 20 geographical phenomena that cannot be modelled satisfactorily by crisp entities as explained in Chapter 2. Suggest alternative data models for these phenomena and discuss the value of the fuzzy set approach in each case.

2. Explain how you would go about measuring the width of geographical boundaries in practice, (*a*) in a landscape, (*b*) in a city.

3. Explore the value of using fuzzy membership functions in sequential data retrieval operations and sieve mapping.

4. Discuss the differences between a *probabilistic* treatment of uncertainty with the *possibilistic* approach of fuzzy sets. When is one approach to be preferred above the other?

Suggestions for further reading

Barrow, J. D. (1992). *Pi in the Sky.* Oxford University Press, Oxford.

Kandel, A. (1986). Fuzzy mathematical techniques with applications. Addison-Wesley, Reading, Mass.

Kauffman, A. (1975). Introduction to the theory of fuzzy subsets. Academic Press, New York.

Klir, G. J., and Folger, T. A. (1988). *Fuzzy Sets, Uncertainty and Information.* Prentice Hall, Englewood Cliffs, NJ.

Kosko, B. (1994). *Fuzzy Thinking: The New Science of Fuzzy Logic.* HarperCollins, London.

Yager, R. R., and Filev, D. P. (1994). *Essentials of Fuzzy Modeling and Control.* John Wiley, New York.

TWELVE

Current Issues and Trends in GIS

GIS is an ever evolving technology although it has reached a certain maturity in recent years. Its development is influenced by many different forces including advances in information technology in general as well as specific changes being initiated by the GIS industry itself. The current areas of negotiation and research into Open GIS, Interoperability, and spatial data infrastructures should lead to the major hurdles of the proprietary nature of systems and data exchange being surmounted. The GIS user community has widened considerably in recent years both in terms of the application fields and the countries in which it is used. Various problems in data handling, data quality, and GIS use still remain.

In this book we have explored many of the conceptual, computational, and mathematical principles of spatial data analysis that form the bases of many GIS. Starting with geographical phenomena and the way people perceive them and build conceptual models of space, we proceeded to demonstrate how the different data models of exact entities and continuous fields are modelled in the computer, and how databases of digital spatial data can be created. We showed how these data can be retrieved and analysed, and how many different methods of spatial analysis provide means to derive new information from raw data, and how logical and numerical models that use these data can provide new understanding, insights, and also the development of new hypotheses for testing. In short, we have seen that the integration of computer technology and geographical data can provide the user with a powerful tool for analysis in the natural and social sciences that greatly extends the capabilities of conventional maps or images.

Since the publication of the first edition of this book (Burrough 1986) the technology (and the associated field of study) has matured considerably, although it is fascinating to find that GIS researchers are still tackling subjects that were major issues ten years ago. For example, the development of efficient storage techniques for huge volumes of geographical data which allow the full dimensional nature of the information to be represented continues to be an important research area. The systems available today are still evolving and the changes in the last few years in data handling and in products are altering the way GIS are used and by whom. At the same time there has been a great increase in the provision and use of geographical data and a growing awareness by organizations of its importance. This has been given impetus by the

demands for an improved understanding of the interactions between people and environment in many investigations, which requires huge amounts of spatial and temporal data to be handled (Burrough 1996*b*, Schell 1995*a*, 1995*b*).

The book has said little about the people who use these systems and the whole industry that has grown up around them. The increasingly complex data handling and analyses provided by GIS means that they are used in policy development, decision-making, and research in a range of political, economic, and academic settings. These systems are being adopted by a broader base of users than ever before and there are large international markets for this technology. GIS are no longer the sole domain of the developed world, with many developing countries using GIS for many applications including cadastral mapping and resource analysis. The use and development of GIS has been supported by an information infrastructure of dedicated books, journals and magazines, conferences and exhibitions, and professional organizations which allow new theoretical, technological, and application developments to be disseminated to a wide audience.

Changes in GIS in the next few years will be governed by (*a*) developments in information technology, (*b*) by various international and industrial standardization agreements, and (*c*) by the expansion of data availability and the provision of associated infrastructures. These will all bring changes in the GIS user community and in the way technology is perceived, and the following is a summary of the main trends.

Changes in technology

Ten years ago a commercial, off-the-shelf GIS was a monolithic all-embracing system with many volumes of manuals that explained the various functions and capabilities; a training period measured in many days, weeks, or even months was required to use it. Today many GIS tools are different and consist of a range of specialized subproducts, such as for cartographic production, database access and retrieval, spreadsheets and spatial data analysis tools, geostatistical interpolation, remote sensing image handling, or computational modelling, which are integrated to a greater or lesser extent. This widespread integration has been made possible by the standardization of data exchange protocols, both between different modules as well as with major data storing and handling tools, such as large relational database management systems. Non-specialists, such as managers and politicians are now in a position to use spatial data in their work without having intermediaries to tackle the complexities of the previous systems. On the other hand, there is a growing distinction between the highly skilled professional who build GIS databases and who are skilled in complex spatial analysis and the majority of users.

GIS have benefited greatly from developments in the IT industry of faster and relatively cheaper computer processors which support their data handling needs. Personal computers and networking have moved GIS from a back-room 'techy' peripheral activity to a desktop application. The current move towards 64-bit computer processors will allow faster data processing and greater memory usage as well as enabling concurrent processing to take place. Whilst these developments will no doubt allow users of GIS to do more complex, data hungry tasks, more quickly, there are likely to be more key changes in systems in the future.

Until recently, most GIS were set up for local, or single main user applications, or for work on a limited project area. As long as the collection, storage, analysis, presentation, and dissemination of geographical information was limited to a single institute, government organization or business unit, differences in hardware, software, and data standards (including data models, database implementation, exchange standards, etc.) were of little importance. The 'best' system for any particular application was determined by examining the functionality, precision, speed, and price-performance relations relative to the user's requirements. GIS were evaluated by 'benchmark tests', as if they were no more than machines on an engineering shop floor. The vendors of GIS protected their interests by not publishing the algorithms, data structures, and procedures so that each system remained largely cut off from every other one (Burrough 1996*b*, Burrough and Masser 1997).

It is no surprise therefore that two topics of current interest to members of the geographical information community are Open GIS and Interoperability (Schell 1995*a*, 1995*b*). They are closely interrelated and refer to the need to create systems which support the efficient description, storage, access, and transfer of geographical data throughout organizations, countries, and even globally. These developments will allow more data to be accessible to a wider audience, promoting more use of geographical information in problem solving and will lead to considerable efficiencies in the data capture process.

The development in ideas of Open GIS, supported in many countries through government action (e.g. the US Federal Data Committee and the National Spatial Data Initiative—NSDI), international collaboration through organizations such as EUROGI (European Umbrella Organization for Geographic Information) and commercial ventures (the Open GIS Consortium) will allow the GI user community to acquire data that may be efficiently exchanged and shared between different software systems. The OGC is concerned with developing technology to exchange data efficiently and to provide links to the larger IT business, so overcoming the current problems of proprietary system databases. Data or GIS purchased through an organization supporting these standards and technology will be readily transferable.

The developments in interoperability aim to remove current constraints on using specific hardware and software platforms for particular data or tasks. For example by setting standards for the database parts of GIS, software from various vendors can be used to query data resident on many different computers. Many commercial systems can now input, or at least display, geographical data in a wide range of formats. The problems of data exchange then pass from the technical domain—can I read your tapes, or can you input the data I am sending over the net?—to the semantic domain. This concerns not the data *per se*, but how they were perceived, recorded, and modelled. A guarantee that one can read in data is no sinecure that they make sense.

The Internet and the World Wide Web will continue to influence GIS use and system developments although their dynamic, uncontrolled nature makes trends somewhat difficult to predict. There is no doubt that both will continue to grow as important sources of information about data and software availability, and for their dissemination. The networks are also likely to become an important GIS 'operating environment' in their own right, with the development of more dynamic interactive tools. Developments in programming languages such as JAVA (of Sun Microsystems), will be useful in this the context as they allow applications (known as Java applets) to be written which are automatically activated by the main web browsing software packages. These applets are easily downloaded from the web server to the local client computer so ensuring faster interaction with the Internet. One of the real benefits of the JAVA language is that the program code is directly portable to various computer platforms running under UNIX, Macintosh, Windows 95, Windows NT, and so on. This will relieve web developers from the need to write different application programs for the various platforms.

Current GIS are most successful at handling static, easily identifiable units in geographical space, and data structuring and handling of more complex data, particularly with a temporal component, continues to challenge GIS developers. In the last few years, database research has been dominated by developments in the object-oriented approach and whilst this approach has benefited the structuring of entity data (especially of easily defined objects, such as property parcels, and utility infrastructure, and offers possibilities for spatio-temporal data handling), there are difficulties in applying it to continuous field data where permanent relations may not exist.

More flexibility and efficiency may be possible through systems that allow multi-scale representation using a nested hierarchical approach in data storage and analysis. These support varying degrees of generalization and abstraction in the data and are likely to gain in popularity as they reflect better our understanding of spatial scales of geographical data. The developments in logic-based databases also offer potential in representing data in a more intuitive manner but this is still an active research area (Worboys 1995).

Three-dimensional GIS are becoming increasingly available (cf. Plates 3.3, 3.4) but the problem of representing the temporal nature of geographical information remains a major hurdle to be overcome. Many current GIS database structures can only represent static 2D information and can only represent temporal changes by either a series of different datasets of an area, or by multi-temporal attributes. This leads to a number of problems including redundancy in the database and clumsiness in integrating the different datasets in any analysis (Langran 1992). It becomes obvious from current developments and reported research that there is still considerable need

for theoretical and technical developments in the modelling of geographical phenomena (Burrough and Frank 1996).

Many human–computer interfaces for GIS now include some means of writing high-level programs (macros) for repetitive tasks. With effort they can sometimes be used for space-time modelling, but are often difficult to use. Recent developments in raster GIS (Mitasova *et al.* 1996, van Deursen and Burrough 1998) are providing powerful, easy to use dynamic modelling tools that are themselves a high-level programming language (cf. Wesseling *et al.* 1996).

Changes in data supply

Data for GIS have been a perennial problem limiting GIS adoption and use although one of the major trends of the last five years has been in the growth in their provision. As the geographical information industry has developed, users realized that cost savings could be made if they could trade spatial data: if another organization has created a digital version of the data you need why spend time and money redoing the work? In practice this data transfer sometimes cost as much or more than the original digitizing because of the lack of data exchange formats available to support this, as discussed in Chapter 4. Until recently, the approach of many governments and other organizations to implement standards for data transfer and quality has been in most cases rather *ad hoc* and this lack of firm policy decisions has led to a poorly controlled, fragmented market in geographical data. Users of data are not always sure of the accuracy or precision of the information and given the burgeoning provision of data from so many sources across the globe on the Internet this has many serious repercussions. Data from different sources which may be combined technically very easily, in terms of different quality, collection standards, georeferencing, etc. may be incompatible leading to spurious analysis results.

Various new initiatives are beginning to address these problems. The developments of Open GIS and their implications for data exchange have already been discussed. In parallel have been various spatial data initiatives, at the national, continental, and global scale, which attempt to ease data availability and integration problems by setting up agreements, guidelines, rules, and policies for the orderly exchange and distribution of geographical data. These initiatives involve establishing large, usually virtual, spatial databases and countries such as the USA and Australia have already made progress with national spatial data infrastructures. There have also been a number of continental and international initiatives (e.g. CORINE 1992, EUROSTAT 1996, UNEP/FAO 1994, Digital Chart of the World 1991) which attempt to collate data from many sources and bring them to common levels of geometry and classification (RIVM 1994). Developments in standards for *metadata* (data about data) that include information about the source, method of recording, resolution, data of collection, data model, and data type are making it easier for users to judge if the data offered are suitable and reliable for their purpose.

These developments are very important if some of the organizational and environmental problems being faced by countries today are to be tackled. Many situations result from actions and accidents in one country which affect the lives and natural resources of persons living in others. The consequences of natural, industrial, economic, and political actions are not restricted by administrative boundaries whether local or international. A particularly poignant demonstration of the international dimension of environmental hazards occurred in Europe in 1986 when the cloud of radioactive waste from the Chernobyl accident roamed the northern hemisphere, ignoring national boundaries and government mandates, to deposit its load on areas far from the source (Karaoglou *et al.* 1996). This, and problems such as the uncontrolled movement of refugees in zones of human conflict need to be addressed through qualitative and quantitative analysis using integrated spatial datasets covering the whole of an affected area rather than just that of one particular country or province.

The difficulties and problems of assembling a uniform geographic database over several administrative areas such as provinces or countries are legion (c.f. Langas 1997). The georeferencing data may refer to different coordinate systems, base levels, or map

projections. Attribute data may have been collected using different sampling methods or size of sample. Laboratory methods, if used to analyse soil, water, air, or biological materials, may be different in different countries or may be subject to varying degrees of error due to differences in local expertise. Some data, in particular social data or land cover data, may be linked to administrative units that are defined differently and have different spatial resolution in different countries (Burrough 1996*b*, Burrough and Masser 1998).

In addition there are problems such as data ownership, legal liability, and standards not to mention the myriad political and institutional issues that need to be resolved. Initiatives to standardize data collection and referencing such as the basic geodetic frameworks for determining geographic location, or elevation data, or thematic data on the location of natural objects, such as rivers, coasts, and lakes, and anthropogenic features such as roads, railways, towns, and cities, often meet with resistance in defending home territories.

These various developments are important in terms of saving money and improving efficiency in an industry in which the costs of collecting, storing, and exchanging data far exceed those of data exploitation (Burrough and Masser 1998). As a consequence markets for geographical information are opened up, bringing sometimes complex and voluminous amounts of data to a wider public and leading to an improvement in decision-making and, importantly, our knowledge and understanding of the spatial distribution of phenomena on the earth's surface.

The semantic problems of data integration and exchange remain more intractable than the technical ones. The problems begin with culture, language, and perception and extends into realms of spatial and temporal analysis. Better, more powerful theoretical concepts are needed than those embedded in the current commercial GIS: this is the domain of GI Science (Burrough 1996*b*, Burrough and Masser 1998).

Changes in the users

GIS users have always ranged across a number of disciplines, reflecting the wide application of geographical data. The systems have been used in studies of the spatial distribution of animals, minerals, and vegetables as well as industrial location, voting patterns, and ancient and modern human habitation. Until recently, most GIS were limited in their scope in terms of both the spatial or study field covered. The systems tended to be used by a number of specialists who were familiar with the software and where responsible for the technology and the database. The generated results for GIS analysis, usually transferred in paper form, were used by a much wider group of people in their work.

The trends in GIS technology towards a series of specialist subproducts in tandem with the increased provision of digital data have brought about a change in users. In some areas, for example spatial modelling, GIS use continues to be a specialist (some might say elitist) activity. However, the development of more specialized modules offering querying and mapping capabilities, support new sets of users who do not necessarily have GIS technical know-how. The easy-to-use interfaces allow users, who have a good understanding of the data, to interact with the system directly, much as they would a spreadsheet. This means that they do not have to communicate their analytical needs to a second person. Intranet developments will also offer organizations an efficient way of providing geographical data and system capabilities to a wide range of people in a user-friendly way. For many users of the current Internet/Intranet data and technology, GIS are essentially digital atlases or gazetteers. These users tend not to collect data and many do not make maps. They are essentially a new type of user—the 'spatial browser'.

One of the marked changes in the user community have been the increasing number of GIS professional people. They offer consultancy and advice to users on purchasing, maintaining, and using systems and are proficient in GIS as a generic idea, and not just as a specific system type. These professionals work for a range of different companies ranging from one-man bands to divisions within major international consultancy organizations. Some GIS professionals are specialized within a particular field such as hydrology

or market research and are proficient at using systems to solve analytical problems in these areas. Various GIS professional organizations have been established in a number of countries and provide a forum for user and system vendor interaction.

The GIS user community is now more worldwide than ever and Internet developments will ensure that trend continues. Their applications are in a wide range of fields and projects, and most interestingly across a variety of cultures. The employment of these systems has met with varying degrees of success with data availability and importantly institutional and societal issues usually being more influential than the technology itself. There have been a number of studies on the impact of GIS on organizations (e.g. Huxhold and Levinsohn 1995) and more recently the larger influences of this type of technology on the thinking of communities have been studied (Bijker *et al.* 1987).

The concepts embodied in GIS reflect the thinking of certain classes of technically oriented people in relatively few parts of the world. The power of the ideas is evident from the speed at which these ideas are assimilated (e.g. the first edition of this book has been translated into Japanese, Chinese, and Thai, and there is an ARC-INFO in every country). Instead of enhancing cultural richness, however, bringing in a system that has been developed largely within the cultural boundaries of an alien mindset often means that the recipients treat the imported system as the only reality and apply it indiscriminately (Burrough 1996*b*). Technology, as it embraces the concepts of progression and development, is often seen as something that must always be followed slavishly.

GIS is an important tool for use in solving problems which require the manipulation and analysis of geographical data. Many of today's ambitions of improving economic and social well-being and environmental improvement go hand in hand, and require a spatial awareness and understanding of processes and form. The Earth Summit at Rio recognized this and many of the actions approved by governments include references to spatial data and geographical information (Thacher 1996) in areas such as soil erosion, marine pollution, biological diversity, and breeding stocks, drinking water supplies, climate change, poverty, urbanization, or demographic change. GIS will continue to provide answers but their successful use will depend on the questions posed, the data available, the institutional and political support necessary to bring about any changes, and of course, on the skills and insights of those who use them.

Suggestions for further reading

Bijker, W. E., Hughes, T. P., and Pinch, T. J. (eds.) (1987). *The Social Construction of Technological Systems: New Directions in the Sociology and History of Technology.* MIT Press, Cambridge.

Burrough, P. A. (1996). A European view of Global Spatial Data Infrastructure. *Proc. Emerging Global Spatial Data Infrastructure.* A Conference held under the patronage of Dr Martin Bangemann, European Commission, EUROGI/Deutscher Dachverband für Geoinformation/Atlantic Institute/Open GIS Consortium/Federal Data Committee/Fédération Internationale des Géomètres (Commission 3), 4–6 Sept. 1996, Königswinter, Germany.

Burrough, P. A., and Masser, F. I. (1998). *European Geographic Information Infrastructures: Opportunities and Pitfalls.* Taylor and Francis, London.

Langran, G. (1992). *Time in Geographic Information Systems.* Taylor & Francis, London.

Mounsey, H. M. (1991). Multisource, multinational environmental GIS: lessons learnt from CORINE. In D. J. Maguire, M. F. Goodchild, and D. W. Rhind (eds.) *Geographical Information System*, ii: *Applications*, Longman Scientific and Technical, Harrow, pp. 185–200.

Schell, D. (1995*a*). What is the meaning of standards consortia? *GIS World* (August 1995), 82

— (1995*b*). Harnessing change. Editorial in *OpenGIS*, Newsletter included in *Geo Info Systems*, May 1995.

Appendix 1. Glossary of Commonly Used GIS Terms

Absolute georeference, the referencing in space of the location of a point using a predefined coordinate system such as latitude and longitude or a national grid.

Abstraction, the division of real world phenomena into individual, distinct items.

Acceptance test, a test for evaluating a newly purchased system's performance and conformity to specifications.

Access time, a measure of the time interval between the instant that data are called from storage and the instant that delivery is complete.

Accuracy, conformance to a recognizable standard; can often mean the number of bits in a computer word available in a given system. The statistical meaning of accuracy is the degree with which an estimated mean differs from the true mean.

Addressability, the number of positions (pixels) in the X and Y axes on a VDU or graphics screen.

Addressable point, a position on a visual display unit (VDU) that can be specified by absolute coordinates.

Algorithm, a set of rules for solving a problem. An algorithm must be specified before the rules can be written in a computer language.

Aliasing. 1. The occurrence of jagged lines on a raster-scan display image when the detail exceeds the resolution of the screen. 2. In Fourier analysis the effect of wavelengths shorter than those sampled by the observation points on the form of the estimated power spectrum.

Altitude matrix, a grid of elevation values.

Alphanumeric code, machine processable letters, numbers, and special characters. Hence alphanumeric screen, alphanumeric keyboard for displaying and entering alphanumeric characters.

American National Standards Institute (ANSI), an association formed by the American Government and Industry to produce and disseminate widely used industrial standards.

American Standard Code for Information Interchange (ASCII), a widely used industry standard code for exchanging alphanumeric codes in terms of bit-signatures.

Analogue. 1. Representation of information by a continuously varying signal (contrast with discretized signals such as digital data). 2. A representation of a physical variable or phenomenon by another variable which displays proportional relationships over a specified range; e.g. using a map to describe an area on the earth's surface.

Anisotropic, an adjective to describe a spatial phenomenon having different physical properties or actions in different directions.

Application, a task addressed by a computer system.

Application program or package, a set of computer programs designed for a specific task.

Arc, a complex line connecting a sequence of coordinate points. Also known as a chain or string.

Archival storage, magnetic and optical media (tapes, removable disks) used to store programs and data outside the normal addressable memory units of the computer.

Area, a fundamental unit of geographic information (a geographic primitive) which is a measure of a particular extent of the earth's surface (*see* point, line, and polygon).

Array, a series of addressable data elements in the form of a grid or matrix.

Array processor, special hardware for high-speed processing of data encoded on a matrix.

Assembler, a computer program that converts programmer-written instructions into computer-executable (binary instructions).

Assembly language, a low-level (primitive) programming language that uses mnemonics rather than English-like statements.

Attribute, non-graphic information associated with a point, line, or area element in a GIS.

Autocorrelation, autocovariance, statistical concepts expressing the degree to which the value of an attribute at spatially adjacent points varies with the distance or time separating the observations.

Automated cartography, the process of drawing maps with the aid of computer driven display devices such as plotters and graphics screens. The term does not imply any information processing.

Background processing/mode. Tasks such as printing are given a lower priority by the computer than those requiring direct user interaction.

Backup, making a copy of a file or a whole disk for safe keeping in case the original is lost or damaged.

BASIC (Beginner's All-purpose Symbolic Instruction Code), a simple, high-level computer programming language, for inexperienced computer users.

Baud rate, a measure of the speed of data transmission between a computer and other devices—equivalent to bits per second.

Benchmark test, a test to evaluate the capabilities of a computer system in terms of the customer's requirements.

Binary arithmetic, the mathematics of calculating in powers of 2.

Binary coded decimal, the expression of each digit of a decimal number in terms of a set of bits.

Binary trees, a data compression and indexing technique which subdivides the spatial or attribute data into a series of levels.

Bit, the smallest unit of information that can be stored and processed in a computer. A bit may have two values—0 or 1; i.e. YES/NO, TRUE/FALSE, ON/OFF.

Bit map, a pattern of bits (i.e. ON/OFF) on a grid stored in memory and used to generate an image on a raster scan display.

Bit plane, a gridded memory in a graphics device used for storing information for display.

Bits per inch (BPI), the density of bits recorded on a magnetic tape; 800, 1600, and 6250 are common standards for old, large tapes.

Block codes, a compact method of storing data in raster databases which simplifies the region into a series of two dimensional blocks of various dimensions, e.g. 2×2, 3×3, etc. The record in the database records the origin, the bottom left, and the side of each square.

Block kriging, the prediction of attribute values for square blocks of land using the kriging methods of geostatistical interpolation.

Boolean operators, these are operators based on logic which are used to query two or more sets of data. The operators allow inclusion, exclusion, intersection, and differences in the data to be determined.

Break points, points on the graph of a continuous variable (e.g. splines) where there is an abrupt change in direction.

Buffering, the creation of a zone of specified width around a point, line, or area. The buffer is a new polygon which is used in queries to determine which entities occur within or outside the defined area.

Bug, an error in a computer program or in a piece of electronics that causes it to function improperly.

Byte, a group of contiguous bits, usually 8, that represent a basic unit of information which is operated on as a unit. The number of bytes is used to measure the capacity of memory and storage units, e.g. 256 kbytes, 300 Mbytes.

C++, a high-level programming language often used to write graphics programs.

Cadastral map, a map showing the precise boundaries and size of land parcels.

Cartography, the art and science of drawing charts and maps.

Cartridge tape, a type of magnetic memory tape enclosed in a plastic cartridge.

Cathode ray tube, the electronic device used in an electronic screen which controls the display of information or graphics.

CD-ROM, a compact disk, used as a *R*ead *O*nly *M*emory device.

Cell, the basic element of spatial information in the raster/grid description of spatial entities (*see* pixel).

Central processing unit (CPU), the part of the computer that controls the whole system.

Chain, *see* arc.

Chain codes, a compact method of storing data in raster databases which simplifies the boundary of a region in terms of a sequential series of north, south, east, or west directional vectors grid, and the number of cells in each direction.

Character, an alphabetical, numerical, or special graphic symbol that is treated as a single unit of data.

Chorochromatic maps, a map in which the area is divided into a series of zones which are each displayed using a single colour or shading.

Choropleth map, a map consisting of a series of single-valued, uniform areas separated by abrupt boundaries.

Classification, the process of assigning items to a group or set according to their attributes.

Clump, the fusion of neighbouring cells which are of the same class into larger units.

Code, a set of specific symbols and rules for representing data and programs so that they can be understood by the computer. *See* ASCII, FORTRAN, PASCAL, etc.

Co-kriging, estimation of a regionalized variable using observations of that variable supplemented by observations of one or more additional variables from within the same geographical area, thereby reducing the estimation variance if the original variable has been undersampled.

Colour display, a computer screen or VDU capable of displaying maps and results in colour.

Command, an instruction sent from the keyboard or other control device to execute a computer program.

Command language, an English-like language for sending commands for complicated program sequences to the computer.

Command Language Interpreter (CLI), a computer program for converting English-language-like commands into instructions for the computer.

Compiler, a computer program that translates a high-level programming language, such as FORTRAN, C++ or PASCAL into machine-readable code.

Composite map, a single map created by joining together several maps that have been digitized separately.

Computer assisted/aided cartography (CAC), the use of computer hardware and specific software for making maps and charts.

Computer graphics, a general term embracing any computing activity that results in graphic images.

Computer word, a set of bits (typically 16 or 32) that occupies a single storage location and is treated by the computer as a unit of information.

Computing environment, the total range of hardware and software facilities provided by a computer and its operating system.

Conceptual model, the abstraction, representation, and ordering of phenomena using the mind.

Conditional simulation, the simulation of a single random function that honours the data values at the sampling points.

Configuration, a particular combination of computer hardware and software for a certain class of application tasks.

Confusion index, a measure of the relative dominance of the membership values assigned to an individual for two or more fuzzy classes.

Connectivity, the linking of different spatial, mostly linear, units into complex chains.

Console, a device that allows the operator to communicate with the computer.

Contiguous, adjacent spatial units touching to form an unbroken chain or surface.

Contour, a line connecting points of equal elevation.

Convolution, the conversion of values from one grid to another which is different in terms of size or orientation.

Cross-hatching, the technique of shading areas on a map with a given pattern of lines or symbols.

Crossover point, the point at which there is a 50 per cent possibility that something belongs to a particular fuzzy class.

Cross validation, a validation method in which observations are dropped one at a time from a sample size n, and n estimates are computed from the remaining $(n-1)$ observations. The statistics of the estimates are used to evaluate the goodness of fit. In geostatistics this is done using the kriging equations to check the variogram with respect to the sample data.

Cursor, a visible symbol guided by the keyboard, a joystick, a tracking ball, or a digitizer, usually in the form of a cross or a blinking symbol, that indicates a position on a computer screen or VDU.

Data analysis models, a series of commands that when combined perform a particular kind of data analysis.

Data link, the communication lines and related hardware and software systems needed to send data between two or more computers over telephone lines, optical fibres, satellite networks, or cables.

Data model, the abstraction and representation of real world phenomena according to a formalized, conceptual schema, which is usually implemented using the geographical primitives of points, lines, and polygons, or discretized continuous fields.

Data record, *see* tuple.

Data structure, the organization of data in ways suitable for computer storage and manipulation.

Data types, the classification of different data according to their characteristics. For example Boolean (0/1), nominal, ordinal, integer, scalar (real), directional, or topological data types according to their function and degree of precision.

Database, a collection of interrelated information, usually stored on some form of mass-storage system such as magnetic tape or disk. A GIS database includes data about the position and the attributes of geographical features that have been coded as points, lines, areas, pixels, or grid cells.

Database Management System (DBMS), a set of computer programs for organizing the information in a database. Typically, a DBMS contains routines for data input, verification, storage, retrieval, and combination.

Debug, to remove errors from a program or from hardware.

Debugger, a program that helps a programmer to remove programming errors.

Delaunay triangulation, the graph obtained by joining pairs of points whose polyhedra are Thiessen (Voronoi/Dirichlet) divisions of the plane.

DEM, *see* Digital Elevation Model.

Device, a piece of equipment external to the computer designed for a specific function such as data input, data storage, or data output.

Differentiable continuous surface, the representation of a continuously varying phenomenon using scalar or integer data so that the rate of change across and within the area may be derived.

Digital, the representation of data in discrete, quantized units or digits.

Digital Elevation Model (DEM), a quantitative model of a part of the earth's surface in digital form. Also digital terrain model (DTM).

Digitize, (noun) a pair of XY coordinates; (verb) to encode map coordinates in digital form.

Digitizer, a device for entering the spatial coordinates of mapped features from a map or document to the computer. A pointer device, a cursor, puck, or mouse is used to locate key points.

Discretization, the process of dividing an area into a series of self-contained units.

Disjunctive kriging, a non-linear distribution-dependent estimator for regionalized variables that do not have simple (Gaussian) distributions. It is the most demanding kriging method in terms of computer resources, mathematical understanding, and stationarity conditions.

Disk, a storage medium consisting of a spinning disk coated with magnetic material for recording digital information.

Diskette, *see* floppy disk.

Dirichlet tessellation, *see* Thiessen polygons.

Distributed processing, the placement of hardware processors where needed, instead of concentrating all computing power in a large central CPU.

Dot-matrix plotter, a plotter of which the printing head consists of many, closely spaced (100–200 per inch) wire points that can write dots on the paper to make a map. Also known as an electrostatic plotter or matrix plotter.

Double precision, typically refers to the use in 32-bit word computers of a double word of 64 bits to represent real numbers to a precision of approximately 16 significant digits.

DPI (dots per inch), a measurement of the density of dots used to print or scan an area with larger values representing more detail and a finer resolution.

Drift, a trend in the data.

Drum scanner, a scanning device for converting maps to digital form, in which the Y-axis movements are governed by the rotation of the drum.

Edit, to remove errors from or to modify a computer file of a program, a digitized map, or a file containing attribute data.

Element, a fundamental geographical unit of information, such as a point, line, area, or pixel. May also be known as an 'entity'.

Ellipsoid, mathematical model for the shape of the earth, taking account of flattening at the poles.
Entities, distinct units of a real world phenomenon.
Exact interpolator, an interpolation method that predicts a value of an attribute at a sample point that is identical to the observed value.
Experimental variogram, an estimate of a (semi-)variogram based on sampling.
Extrapolation, the estimation of the values of an attribute at unsampled points outside an area covered by existing measurements.
Feature planes, a series of separate different classes of phenomena which are often used as the basis of the different overlays.
Field. 1. A type or class of data. 2. Within a database, a set of records containing information (cf. tuple).
File, a collection of related information in a computer that can be accessed by a unique name. Files may be stored on tapes or disks.
Filter, in raster graphics, a mathematically defined operation for removing long-range (high-pass) or short-range (low-pass) variation. Used for removing unwanted components from a signal or spatial pattern.
Finite difference modelling, a numerical modelling technique used with data held in regular grid form in which algebraic equations are used to solve changes in a variable at each location.
Finite element modelling, a numerical modelling technique used with data held in irregular grid (usually triangular) form in which algebraic equations are used to solve changes in a variable at each location.
Flatbed plotter, a device for drawing maps whereby the information is drawn by the plotting head being moved in both the X and Y directions over a flat, fixed surface. Draws with a pen, light beam, or scribing device.
Floating point, a technique for representing numbers without using a fixed-position decimal point in order to improve the calculating capability of the CPU for arithmetic with real numbers.
Floating point board, a printed circuit board placed in the CPU in order to speed up arithmetic operations for real numbers. (The alternative is to use special software, which is usually much slower.)
Floppy disk, a cheap, low-capacity storage medium, usually measuring 3.5 inches in diameter, capable of storing up to 2.0 Mbytes of data. Also known as a diskette or floppy.
Font, symbolism used for drawing a line, or the name of a typeface used for displaying text.
Format, the way in which data are systematically arranged for transmission between computers, or between a computer and a device. Standard format systems are used for many purposes.
FORTRAN (FORmula TRANslation), a high-level programming language, much used in computer graphics. Recent improvements, embodied in FORTRAN 77, have made structured programming and interactive data input much easier.
Fourier analysis, a method of dissociating time series or spatial data into sets of sine and cosine waves.
Fractal, an object having a fractional dimension; one which has variation that is self-similar at all scales, in which the final level of detail is never reached and never can be reached by increasing the scale at which observations are made.
Fuzzy set, a set of objects in which the membership of the set is expressed in terms of a continuous membership function having values between 0 and 1. Unlike Boolean sets, fuzzy sets can overlap and an individual can be a member of the overlapping sets to different degrees.
Gap, the distance between two graphic entities (usually lines) on a digitized map. Gaps may arise through errors made while digitizing or scanning the lines on a map.
Generalization, the process of reducing detail on a map as a consequence of reducing the map scale. The process can be semi-automated for certain kinds of data, such as topographical features, but requires more insight for thematic maps.
Geocoding, the activity of defining the position of geographical objects relative to a standard reference grid.
Geodetical surveying, the determination of the position of points on the earth's surface accounting for its curvature, rotation, and gravitational field.
Geographical data, data that record the location and a value characterizing the phenomenon.
Geographical data model, formalized schema for representing data which has both location and characteristic.
Geographical information system, a set of computer tools for collecting, storing, retrieving at will, transforming, and displaying spatial data from the real for a particular set of purposes.
Geographical primitives, the smallest units of spatial information: in vector form these are points, lines, and areas (polygons); in raster form they are pixels (2D) and voxels (3D).
GPS (Global Positioning System), a set of satellites in geostationary earth orbits used to help determine geographic location anywhere on the earth by means of portable electronic receivers.
Graphics tablet, a small digitizer (usually measuring 30 × 30 cm) used for interactive work.
Grey scales, levels of brightness (or darkness for displaying information on monochrome display devices.
Grid. 1. A set of regularly spaced sample points. 2. A tessellation by squares. 3. In cartography, an exact set of reference lines over the earth's surface. 4. In utility mapping, the distribution network of the utility resources, e.g. electricity or telephone lines.
Grid map, a map in which the information is carried in the form of regular squares. Also called a raster.
Hardcopy, a copy of a graphics or map image on paper or other stable material.
Hard data, data obtained by direct measurement.

Hardware, the physical components of a GIS—the computer, plotters, printers, VDUs, and so on.

Hexadecimal system, the representation of numbers and letters using base 16 alphanumeric values.

Hidden line removal, a technique in 3D perspective graphics for suppressing the appearance of lines that ordinarily would be obscured from view.

Hierarchical database structure, a method of arranging computer files or other information so that the units of data storage are connected by a hierarchically defined pathway. From above to below, relations are one-to-many.

High-level language, a computer programming language using command statements, symbols, and words that resemble English-language statements. Examples are FORTRAN, PASCAL, C^{++}, PL/1, COBOL, BASIC.

Histogram, a diagram showing the number of samples that fall in each contiguously defined size class of the attribute studied.

Hole effect, a condition in which the variogram does not increase monotonically beyond the range. The cause may be real or pseudo periodicities in the sample data.

Host computer, the primary or controlling computer in a data network.

Hypsometry, the measurement of the elevation of the earth's surface with respect to sea level.

Indexed files, files of data records in which pointers, based on a particular ordering such as alphabetic, are used to quicken accessing and searching instructions.

Indicator kriging, a kriging interpolation method which is non-linear and in which the original data are transformed from a continuous to a binary scale.

Inexact interpolator, interpolation methods that provide estimates at data locations that are not necessarily the same as the original measurements.

Input, (noun) the data entered to a computer system; (verb) the process of entering data.

Input device, a hardware component for data entry; *see* digitizer, keyboard, scanner, tape drive.

Integer, a number without a decimal component; a means of handling such numbers in the computer which requires less space and proceeds more quickly than with numbers having information after the decimal point (real numbers).

Interactive, a GIS system in which the operator can initiate or modify program execution via an input device and can receive information from the computer about the progress of the job.

Interface, a hardware and software link that allows two computer systems or a computer and its peripherals to be connected together for data communication.

Interpolation, the estimation the values of an attribute at unsampled points from measurements made at surrounding sites.

Intersection. 1. Geometric: the crossing of lines or polygons to form new units. 2. Logic: the combination of data from two Boolean sets using the AND operator.

Intrinsic hypothesis, a form of spatial stationarity less restrictive than second order stationarity, in which the stationarity requirements are confined to the first differences and not the underlying regionalized variable. The intrinsic hypothesis is useful for modelling regionalized variables in which the form of the variogram is a function of domain size.

Isoline, a line which joins points of equal value.

Isopleth map, a map displaying the distribution of an attribute in terms of lines connecting points of equal value; *see* contour, *contrast with* choropleth map.

Isotropic, an adjective to describe something that has the same physical properties or actions in all directions.

Join. 1. (verb) to connect two or more separately digitized maps; 2. (noun) the junction between two such maps, sometimes visible as a result of imperfections in the data.

Justification (right, left, or centre), the relative position of a text string or symbol on the map to the location at which it has been digitized.

Key file, in some CAD/CAM systems, a file containing the codes defining the operation of certain keyboard functions, or menu commands. In DBMS, a file containing information about search paths or indexes used to access data.

Keyboard, the device used for typing information into the computer.

Kriging, name (after D. G. Krige) for a suite of interpolation techniques that use regionalized variable theory to incorporate information about the stochastic aspects of spatial variation when estimating interpolation weights.

Kriging variance, a measure of the uncertainty of estimation for values predicted by kriging.

LANDSAT, a series of earth resource scanning satellites launched by the United States of America.

Laser printer, a printer in which the information is written onto light-sensitive drum or material using a laser.

Layer, a logical separation of mapped information according to theme. Many geographic information systems and CAD/CAM systems allow the user to choose and work on a single layer or any combination of layers at a time.

Legend, the part of a map explaining the meaning of the symbols used to code the depicted geographical elements.

Library, a collection of standard, often used computer subroutines, or symbols in digital form.

Light pen, a hand-held photosensitive interactive device for identifying elements displayed on a refreshed computer screen.

Line, one of the basic geographical primitives, defined by at least two pairs of XY coordinates.

Line printer, a printer that prints a line of characters at a time.

Linear interpolator, describes a method whereby the weights assigned to different data points are computed using a linear function of distance between sets of data points and the point to be predicted.

Local drain direction (ldd), the direction of steepest downhill slope as determined from a gridded DEM.

Lookup table, an array of data values that can be quickly accessed by a computer program to convert data from one form to another, e.g. from attribute values to colours.

Machine language, instructions coded so that the computer can recognize and execute them.

Macro, a text file containing a series of frequently used operations that can be executed by a single command. Can also refer to a simple high-level programming language with which the user can manipulate the commands in a GIS.

Magnetic media, tape or disks coated with a magnetic surface used for storing electronic data.

Mainframe, a large computer supporting many users.

Map. 1. A hand-drawn or printed document describing the spatial distribution of geographical features in terms of a recognizable and agreed symbolism. 2. A collection of digital information about a part of the earth's surface.

MAP (Map Analysis Package), a computer program written by C. D. Tomlin for analysing spatial data coded in the form of grid cells.

Mapping unit, a set of areas drawn on a map to represent a well-defined feature or set of features. Mapping units are described by the map legend.

Map projection, the basic system of coordinates used to describe the spatial distribution of elements in a GIS.

Mass storage system, auxiliary, large-capacity memory for storing large amounts of data. Usually optical or magnetic disk or tape.

Maximum likelihood, a method embodying probability theory for fitting a mathematical model to a set of data.

Mean, the average, or most likely value.

Menu, a list of available options displayed on the computer screen that the user can choose from by using the keyboard or a device such as a mouse.

Metadata, information about the provenance, resolution, availability, age, ownership, price, copyright, and other matters concerning digital spatial data that is easily available to potential users.

Minicomputer, a medium-sized, general purpose single processor computer often used to control GIS (*see* workstation).

Model. 1. A representation of attributes or features of the earth's surface in a digital database. 2. A set of algorithms written in computer code that describe a given physical process or natural phenomenon of the earth's surface. 3. A function fitted to an experimental variogram derived from sample data. 4. A statistical distribution or a conceptualization of spatial variation.

Modem (MOdulator-DEModulator), a device for the interconversion of digital and analogue signals to allow data transmission over telephone lines.

Module, a separate and distinct piece of hardware or software that can be connected with other modules to form a system.

Morton ordering, a technique for reducing the geographical referencing of grid data to one dimension by following a set 'Z' shape directional pattern through the cells.

Mouse, a hand-steered device for entering data or commands to a computer.

Nested sampling, the measurement of data at a series of points whose locations are hierarchically structured.

Network 1. Two or more interconnected computer systems for the implementation of specific functions. 2. A set of interconnected lines or arcs.

Network database structure, a method of arranging data in a database so that explicit connections and relations are defined by links or pointers of a many-to-many type.

Node, the point at which arcs (lines, chains, strings in a polygon network) are joined. Nodes carry information about the topology of the polygons.

Noise, irregular variations or error, usually short range, that cannot be easily explained or associated with major mapped features or process.

Non-differentiable continuous surface, representation of a continuously varying phenomenon using binary, nominal, or ordinal data types that do not support the calculation of the rate of change across and within an area.

Non-linear kriging, *see* indicator kriging *and* disjunctive kriging.

Non-removable storage, computer data storage which cannot be moved easily and includes hard disks.

Non-transitive variogram, a variogram in which the sill is not reached within the domain of interest (*see* intrinsic hypothesis).

Normalization, methods that are used to reduce redundancy and improve efficiency in a database.

Nugget, in kriging and variogram modelling, that part of the variance of a regionalized variable that has no spatial component (variation due to measurement errors and short-range spatial variation at distances within the smallest inter-sample spacing).

Numerical taxonomy, quantitative methods for classifying data using computed estimates of similarity.

Object code, a computer program that has been translated into machine readable code by a compiler.

Object-oriented database structure, the organization of data within a database defined by a series of predefined objects and their properties and behavioural characteristics.

ODYSSEY, computer program developed at the Laboratory for Computer Graphics, Harvard, for overlaying polygon networks.

Operating system (O/S), the control program that coordinates all the activities of a computer system.

Optimal estimator, an estimator for minimizing the value of a given criterion function; in kriging this is the estimation variance.

Ordered sequential files, a file of data records which are in sequence according to some structuring method such as the alphabet.

Ordinary kriging, a method for interpolating data values from sample data using regionalized variable theory in which the prediction weights are derived from a fitted variogram model.

Orthophotos, a scale-correct photomap created by geometrically correcting aerial photographs or satellite images.

Output, the results of processing data in a GIS; maps, tables, screen images, tape files.

Overlay. 1. (verb) the process of stacking digital representations of various spatial data on top of each other so that each position in the area covered can be analysed in terms of these data; 2. (noun) a data plane containing a related set of geographic data in digital form.

Package, a set of computer programs that can be used for a particular generalized class of applications.

Paint, to fill in an area with a given symbolism on a raster display device (*see* cross-hatching).

PASCAL, a high-level computer programming language.

Peano-Hilton ordering, a technique for reducing the geographical referencing of grid data to one dimension by following a recursive route through the cells.

Pen plotter, a device for drawing maps and figures using a computer steered pen.

Performance, the degree to which a device or system fulfils its specifications.

Peripheral, a hardware device that is not part of the central computer.

Photogrammetry, a series of techniques for measuring position and altitude from aerial photographs or images using a stereoscope or stereoplotter.

Photomosaic, a collection of aerial photographs which are joined to form a contiguous view of an area.

Pit, a depression on the surface of a digital elevation model that may represent a real feature or which may be an artefact of the gridding.

Pixel, contraction of picture element; smallest unit of information in a grid cell map or scanner image.

Plotter, any device for drawing maps and figures.

Polygon, a multi-sided figure representing an area on a map; a geographic primitive.

Polygon overlay and intersection, the creation of new polygons (entities) by the process of overlaying and intersecting the boundaries from two or more vector representations of area entities.

Polynomial, an expression having a finite number of terms of the form $ax + bx5 + \ldots nx$.

Precision. 1. Degree of accuracy of numerical representation; generally refers to the number of significant digits of information to the right of the decimal point. 2. Statistical; the degree of variation about the mean.

Principal component analysis (PCA), a method of analysing multivariate data in order to express their variation in terms of a minimum number of principal components or linear combinations of the original, partially correlated variables.

Prism, 1. A polygonal solid; a polyhedron having parallel, polygonal, and congruent bases and sides that are parallelograms. 2. Sometimes used in GIS to indicate a 3D solid body delineated by irregular polygonal faces.

Probability, the chance of an event or occurrence.

Probability distribution function, a real-valued function (in the range 0, 1) whose integral over a set gives the probability of a random variable having a value within the set.

Program, a set of instructions directing the computer to perform a task.

Proximity, the closeness of one item to another.

Puck, a hand-held device for entering data from a digitizer which usually has a window with accurately engraved cross-hairs, and several buttons for entering associated data.

Quadrant, a quarter of a circle measured in units of 90 degrees.

Quadratic polynomial, one in which the highest degree of terms is 2.

Quadtree, a data structure for thematic information in a raster database that seeks to minimize data storage.

Range. 1. In arithmetic, the difference between the largest and smallest values in a set. 2. In geostatistics, the distance at which a transitive variogram ceases to increase monotonically.

Raster, a regular grid of cells covering an area.

Raster data structure, a database containing all mapped, spatial information in the form of regular grid cells.

Raster display, a device for displaying information in the form of pixels on a computer screen or VDU.

Raster map, a map encoded in the form of a regular array of cells.

Rasterization, the process of converting an image of lines and polygons from vector representation to a gridded representation.

Raster-to-vector conversion, *see* vectorization.

Real data, numbers that have both an integer and a decimal component (scalars).

Real time, tasks or functions executed so rapidly that the user gets an impression of continuous visual feedback.

Realization, an equi-probable result of stochastic simulation based on a known probability distribution function.

Record, a set of attributes relating to a geographical entity; a set of related, contiguous data in a computer file.

Redundancy, the inclusion of data in a database that contribute little to the information content.

Region, a set of loci or points having a certain value of an attribute in common.

Regionalized variable, a single-valued function defined over a metric space (a set of coordinates) that represents the variation of natural phenomena that are too irregular at the scale of interest to be modelled analytically.

Relational database structure, a method of structuring data in the form of sets of records or tuples so that relations between different entities and attributes can be used for data access and transformation.

Relative georeferencing, the referencing in space of the location of a point to a local base station rather than to a global grid.

Removable storage, computer data magnetic or optical storage media which are portable and include examples such as floppy disks, DAT tapes, and CD ROMs.

Resampling, technique for transforming a raster image from one particular scale and projection to another.

Resolution, the smallest spacing between two displayed or processed elements; the smallest size of feature that can be mapped or sampled.

Response time, the time that elapses between sending a command to the computer and the receipt of the results at the workstation.

R-trees, a spatial indexing technique which groups entities according to their proximity by using minimum bounding rectangles. Hierarchies of rectangles may be established. When querying the database any search is directed to the rectangle and any subsequent lower-level ones which contain the item of interest.

Run-length codes, a compact method of storing data in raster databases which simplifies the grid on a row-by-row basis by coding the start and end values of contiguous cells for each class.

Sampling, the technique of obtaining a series of measurements to obtain a satisfactory representation of the real world phenomenon being studied.

Scale, the relation between the size of an object on a map and its size in the real world.

Scanner, a device for converting images from maps, photographs, or from part of the real world into digital form. The scanning head is made up of a light or other energy source and a sensing device which records digital values of light reflected back from the surface.

Scenario, a result of a numerical simulation model in which certain input data may be given values to represent conditions not yet observed. Scenarios are often used to compare forecasts of how landscape changes may turn out.

Semivariogram. 1. Given two locations $\boldsymbol{x}$ and $(\boldsymbol{x} + \boldsymbol{h})$, a measure of one-half of the mean square differences (the semivariance) produced by assigning the value $z(\boldsymbol{x} + \boldsymbol{h})$ to the value $z(\boldsymbol{x})$, where $\boldsymbol{h}$ (known as the lag) is the inter-sample distance. 2. A graph of semivariance versus lag $\boldsymbol{h}$.

Semivariogram model, one of a series of mathematical functions that are permitted for fitting the points on an experimental variogram (linear, spherical, exponential, Gaussian, etc.).

Sill, the maximum level of semivariance reached by a transitive semivariogram.

Simple kriging, an interpolation technique in which the prediction of values is based on a generalized linear regression under the assumption of second order stationarity and a known mean.

Simulation, using the digital model of the landscape in a GIS for studying the possible outcome of various processes expressed in the form of mathematical models.

Sink, *see* pit.

Sliver, a narrow gap between two lines created erroneously by digitizing or by the vectorization software of a scanner.

Smoothing, a set of procedures for removing short-range, erratic variation from lines, surfaces, or data series.

Smoothing spline, a method of fitting a smooth polynomial function through erratic data to capture the long-range variation and to suppress local components.

Spatial data model, *see* geographical data model.

Spheroid, a geometric representation of the shape of the earth.

Soft data, data obtained by inspection, intuition, or from other parties. Not measured directly and therefore often judged to be less reliable than hard data.

Software, general name for computer programs and programming languages.

Source code, a computer program that has been written in an English-language-like computer language. It must be compiled to yield the object code before it can be run on the computer.

Spike. 1. An overshoot line created erroneously by a scanner and its raster-vector software. 2. An anomalous data point that protrudes above or below an interpolated surface representing the distribution of the value of an attribute over an area.

Spline, a polynomial curve or surface used to represent spatial variation smoothly.

Stationarity, a statistical name for expressing degrees of invariance in the properties of random functions; it refers to the statistical model, and not to the data. Most commonly used to indicate invariance in the mean and variance, but also in the variance of first differences (*see* intrinsic hypothesis).

Statistical moments. First order is the mean; second order are the variance, the covariance, and the semivariance.

Stereo plotter, a device for extracting information about the elevation of landform from stereoscopic aerial photographs. The results are sets of X, Y, and Z coordinates.

Stochastic imaging, *see* conditional simulation.

Stochastic simulation, simulation using a probabilistic model to generate a range of allowable data values.

Storage, the parts of the computer system used for storing data and programs (*see* archival storage, magnetic media).

Stratified kriging, interpolation by any kriging method within a set of strata or divisions of the land into different classes.

Structured Query Language (SQL), a standard language for interrogating and managing relational databases.

Support, this is the term used in geostatistics for the area or volume of the sample on which measurements are made (e.g. a volume of soil or water, or a pixel in a remotely sensed image).

SYMAP (SYnagraphic MAPping program), the original grid-cell mapping program developed by Howard T. Fisher at Harvard.

Syntax, a set of rules governing the way statements can be used in a computer language.

Tablet, a small digitizer used for interactive work on a graphics workstation.

Tape drive, a device for reading and writing digital information from and to magnetic tapes.

Terminal, a device, usually including a computer screen or VDU and a keyboard for communicating with the computer.

Tessellation, the process of dividing an area into smaller, contiguous tiles with no gaps in between them.

Text editor, a program for creating and modifying text files.

Thematic map, a map displaying selected kinds of information relating to specific themes, such as soil, land use, population density, suitability for arable crops, and so on. Many thematic maps are also choropleth maps, but when the attribute is modelled by a continuous field, representation by isolines or colour scales is more appropriate.

Thiessen polygons, a tessellation of the plane such that any given location is assigned to a tile according to the minimum distance between it and a single, previously sampled point. Also known as Dirichlet tessellation or Voronoi polygons.

Tile, a part of the database in a GIS representing a contiguous part of the earth's surface. By splitting a study area into tiles, considerable savings in access times and improvements in system performance can be achieved.

Tiling, the creation of a seamless spatial coverage by joining contiguous areas (tiles) together.

Topographical map, a map showing the surface features of the earth's surface (contours, roads, rivers, houses, etc.) in great accuracy and detail relative to the map scale used.

Topology, a term used to refer to the continuity of space and spatial properties, such as connectivity, that are unaffected by continuous distortion. In the representation of vector entities, connectivity is defined *explicitly* by a directed pointer between records describing things that are somehow linked in space (for example a junction between two roads). In regular and irregular tessellations of continuous surfaces (e.g. grids) the topological property of connectivity between different locations may only be *implicitly* defined by the spatial rate of change of attribute values over the grid. The topology (connectivity) of gridded surfaces can be revealed by computing first, second, or higher order derivatives of the surface (see Chapter 8).

Transect, a set of sampling points arranged along a straight line.

Transfer function. 1. A numerical method of transferring spatial data from one projection to another. 2. A numerical model for computing new attribute values from existing data using regression models or other algorithms.

Transform, the process of changing the scale, projection, or orientation of a mapped image.

Transitive variogram, a semivariogram having a range and a sill.

Trend surface analysis, methods for exploring the functional relationship between attributes and the geographical coordinates of the sample points.

Triangular Irregular Network (TIN), a vector data structure for representing geographical information that is modelled as a continuous field (usually elevation) which uses tessellated triangles of irregular shape (*see* Delaunay triangulation).

Tuple, a set of values of attributes pertaining to a given item in a database. Also known as a record.

Turnkey system, a GIS or CAD/CAM system of hardware and software that is designed, supplied, and supported by a single manufacturer ready for use for a given class of work.

Union. 1. Databases: the joining of two or more datasets together. 2. Boolean logic: the joining of two sets using the 'OR' operator

Universal kriging, simple kriging of the residuals of a regionalized variable after systematic variation has been modelled by a drift or trend surface.

UNIX, a computer operating system.

Upstream element map, a map showing the cumulative catchment areas for each cell according to the topology of the local drain direction map.

Utility, a term for system capabilities and features for processing data.

Utility mapping, a special class of GIS applications for managing information about public infrastructure such as water pipes, sewerage, telephone, electricity, and gas networks.

Variogram, common term for semivariogram.

Vector. 1. Physics: a quantity having both magnitude and direction. 2. GIS: the representation of spatial data by points, lines, and polygons.

Vector data structure, a means of coding and storing point, line, and areal information in the form of units of data expressing magnitude, direction, and connectivity.

Vectorization, the conversion of point, line, and area data from a grid to a vector representation.

Vector-to-raster-conversion, *see* rasterization.

Viewshed, those parts of the landscape that can be seen from a particular point.

Visual display unit (VDU), a computer screen used for graphical display.

Voronoi polygon, *see* Thiessen polygons.

Voxels, three-dimensional, cubic units of space.

Weighted moving average, value of an attribute computed for a given point as an average of the values at surrounding data points taking account of their distance or importance.

Window, an area (usually square) that is used to capture data needed to compute derived attributes from an original map.

Windows 95, **Windows NT**, operating systems used with personal computers and workstations.

Word, a set of bits (typically 16 or 32) that occupies a single storage location and is treated by the computer as a unit of information.

Workstation, a minicomputer or high-level personal computer used for local computations; it is often connected to other computers by a network. The operating system used for workstations is often Unix or Windows NT.

Zero, the origin of all coordinates defined in an absolute system. Where X, Y, and Z axes intersect.

Zoom, a capability for proportionately enlarging or reducing the scale of a figure or maps displayed on a computer screen or VDU.

Appendix 2. A Selection of World Wide Web Geography and GIS Servers

This appendix lists sources of free or cheap software which can be used to carry out many or most of the operations reviewed in the text, so that readers can create their own coursework based on the software of their choice. It also gives a very select list of World Wide Web sites where data and information are readily available. In several cases, the sites given are themselves lists of tens or even hundreds of GIS, GPS, Remote Sensing, or Digital Cartography sites, so the reader should be able to find a very wide variety of material very easily indeed.

Sources of software that can be used to support the material in this book

GIS and Geostatistics

Utrecht University: PCRaster
http://www.frw.ruu.nl/pcraster.html
Source of downloadable software for raster mapping and dynamic modelling via
ftp://pop.frw.ruu.nl/pcraster/version2/
Gstat (Geostatistics—E. Pebesma). Information, dynamic demonstrations, software and manuals can be obtained from http://www.frw.uva.nl/~pebesma/gstat/

GRASS
http://www.cecer.army.mil:80/welcome.html — CERL/GRASS welcome file.
http://www.cecer.army.mil/grass/GRASS.main.html — GRASS GIS home page.
ftp://moon.cecer.army.mil/grass — GRASS GIS source and data for Spearfish (South Dakota, US).
ftp://topquark.cecer.army.mil/pub/grass — GRASS GIS source for LINUX.

IDRISI
ftp://midget.towson.edu/idrisi — IDRISI-L FTP site (donated data, modules and more).

Intergraph
http://www.ingr.com/iss/products/mapping/ — Intergraph Corporation.

ILWIS
http://www.itc.nl/homepage.html — ITC—International Institute for Aerospace Survey and Earth Sciences, NL. (Ilwis)

Variowin (Y. Pannetier).
http://www.springer-ny.com/supplements/variowin.html

Global Positioning Systems (GPS)

gopher://unbmvs1.csd.unb.ca:1570/1EXEC%3aCANSPACE>CANSPACE—Canadian Space Geodesy Forum archive (lots of GPS info).
http://www.ggrweb.com — GeoWeb—Online resources for GIS/GPS/RS.
http://ageninfo.tamu.edu/geoscience.html — GIS/Remote Sensing/GPS/Geosciences top level page at Texas A&M
telnet://fedworld.gov — GPS Information Center, U. S. Coast Guard (login: select database 34).
ftp://unbmvs1.csd.unb.ca/PUB.CANSPACE.GPS.INFO.SOURCES>GPS Information Sources (maintained by Richard Langley).
finger gps@geomac.se.unb.ca — GPS satellite current constellation status.
http://sideshow.jpl.nasa.gov/mbh/series.html — NASA—Global GPS Time Series.

General lists and sources of information on GIS, digital data, digital cartography, remote sensing and other sites

The University of Utrecht
http://www.frw.ruu.nl/cgi-bin/nph-count?width=5&link=/nicegeo.html
Oddens Bookmarks (http://kartoserver.frw.ruu.nl/HTML/oddens.html — Utrecht University: maps, atlases, cartography and related matters (maintained by R. P. Oddens).

The European Umbrella Organisation for Geographical Information
http://www.frw.ruu.nl/eurogi/eurogi.html" — EUROGI—European Umbrella Organisation for Geographical Information

The European Science Foundation GISDATA programme
http://www.shef.ac.uk/uni/academic/D-H/gis/gisdata/html

The US National Center for Geographic Information and Analysis (NCGIA):
ftp://ncgia.ucsb.edu/pub — NCGIA FTP server at U. C. Santa Barbara (UCSB).
http://www.ncgia.ucsb.edu/ — NCGIA WWW server at U. C. Santa Barbara (UCSB).

http://zia.geog.buffalo.edu/ — NCGIA WWW server SUNY Buffalo.
The Alexandria Digital Map Library.

The US Geological Survey

http://www.usgs.gov/research/gis/title.html — USGS GIS home page.
http://www.usgs.gov/data/index.html — USGS Data Available Online.
http://www.usgs.gov/fgdc-catalog/products/Digital_Airborne_&_Satellite_Imagery.html — USGS Digital Airborne and Satellite Imagery.
ftp://edcftp.cr.usgs.gov/pub/data/DEM/250 — USGS GISLab FTP site — 1 : 250 000 DEM depot.
http://sun.cr.usgs.gov/gis/hyper/glossary/ — USGS GIS glossary
http://www.usgs.gov/network/science/earth/gis.html — USGS GIS
http://h2o.er.usgs.gov/nsdi/dcw/dcwindex.html — USGS DCW—Digital Chart of the World (US data).
ftp://sdts.er.usgs.gov/pub/dlge — USGS DLG-E Information.
telnet://glis.cr.usgs.gov:guest — USGS Global Land Information System (GLIS)—LANDSAT Archive.
ftp://edcftp.cr.usgs.gov/pub/data — USGS/EROS FTP site—DEM and DLG file depot.
ftp://waisqvarsa.er.usgs.gov/wais/docs — USGS FGDC (US) —Spatial Metadata Standard text source.
ftp://isdres.er.usgs.gov/.USGS.GCTP — USGS GCTP—General Cartographic Transformation Package.
ftp://alum.wr.usgs.gov/pub/map — USGS Geologic fault database for the U.S.
http://www.usgs.gov/fgdc-catalog/title.html — USGS FGDC—US Manual of Federal Geographic Data Products.
telnet://xglis.cr.usgs.gov:5060/ — USGS GLIS—EROS Data Center Global Land Information System (X Windows only).
http://kai.er.usgs.gov — USGS PROJ—projection conversion program source.
ftp://sdts.er.usgs.gov/pub/sdts/>
USGS SLIS—US Spatial Data Transfer Standard (FIPS 173).

The University of Sheffield, UK

http://www.shef.ac.uk/ — University of Sheffield:
http://www.shef.ac.uk/uni/projects/sc — Society of Cartographers,
http://www.shef.ac.uk/uni/academic/D-H/dpsc — the Centre for Development Studies, http://www.shef.ac.uk/uni/academic/D-H/eoc — SCEOS: Sheffield Centre for Earth Observation Science,
http://www.shef.ac.uk/uni/academic/I-M/idry — Centre for International Drylands Research,
http://www.shef.ac.uk/uni/academic/I-M/merc — Migration and Ethnic Centre
http://www.shef.ac.uk/uni/academic/N-Q/perc — PERC: the Political Economy Research Centre.

Appendix 3. Data Sets Used in the Examples

This appendix provides information on some of the data sets used in Chapter 5, 6, 8, and 10, where this is not given in the text.

The digital elevation models used for slope mapping and insolation in Chapter 8

The small DEM has a cell size of 30 × 30 m, an area of 6.04 km^2, a minimum elevation of 175 m, and a maximum elevation of 560 m: the large DEM has a cell size of 60 × 60 m, an area of 175.5 km^2, a minimum elevation of 160 m, and a maximum elevation of 677.5 m.

The Catsop digital elevation model used in Chapter 10

This has a cell size of 10 × 10 m, an area of 41.6 ha, a minimum elevation of 80 m, and a maximum elevation of 112 m. Further details are given in de Roo *et al.* (1989).

Data: Guyana rainforest

The soil map data come from the Tropenbos Ecological Reserve at Mabura and were obtained by conventional soil survey using aerial photographs and field survey (Jetten 1994). The forest inventory of tree species was carried out in the Waraputra Compartment of the Demerara Timbers Limited concession 20 km east of the Tropenbos Reserve (Jetten 1994, Ter Steeg *et al.* 1993).

The soil pollution data set used in Chapters 5 and 6

Note: all coordinates and elevations refer to local, not absolute positions.

Soil samples were collected as bulked samples within a radius of 5 m.

Flood frequency classes and soil types were obtained from Dutch Ministry of Public Works surveys. For more details see Burrough *et al.* (1996). The following file is in GEOEAS format.

Maas pollution data 98 sites sorted on Y, X

11	number of attributes
easting	m—local coordinates
northing	m—local coordinates
elevation	m—above local reference level
D-river	m—from edge of main channel
Cd	ppm
Cu	ppm
Pb	ppm
Zn	ppm
LOI	percent organic matter loss on ignition
Fldf	flood frequency class
Soil	soil type

1637	2651	8.06	10	8.2	47	191	812	11.1	1	1
1894	2630	7.51	170	2.4	32	102	298	1.4	1	2
2110	2630	6.98	340	3.0	32	97	321	1.6	1	2
2356	2618	7.80	550	1.6	27	82	213	3.1	1	2
1755	2599	7.94	100	4.2	51	281	746	5.1	1	2
1503	2597	7.98	10	17.0	128	405	1548	12.3	1	1
1665	2597	8.80	60	2.4	47	297	832	10.0	2	1
1373	2555	7.90	10	12.0	117	654	1839	16.5	1	1
1226	2517	7.74	10	9.4	104	482	1528	13.9	1	1
1500	2513	6.68	70	10.9	90	541	1571	10.2	1	1
1961	2493	8.18	320	1.2	21	48	167	1.0	2	0
1851	2475	8.69	260	1.7	22	65	176	1.0	2	0
1087	2461	7.55	20	8.2	76	276	933	8.1	1	1

1692	2457	6.36	200	4.3	50	294	746	5.3	1	2
1249	2442	8.54	70	1.2	30	244	703	8.3	2	1
1302	2413	6.74	140	3.5	34	207	550	5.8	1	1
1507	2401	9.52	190	0.0	31	96	262	5.9	2	1
1004	2359	7.40	20	7.3	80	310	1190	12.0	1	1
1192	2335	7.76	200	2.6	36	180	432	3.1	1	1
1810	2331	9.28	360	1.3	21	62	258	2.0	1	2
1683	2314	7.78	320	3.1	38	211	464	4.5	1	2
1042	2255	8.77	70	0.0	25	94	253	8.1	2	1
1862	2247	9.81	450	0.0	20	56	142	5.0	2	2
926	2236	7.44	10	9.4	78	210	907	14.1	1	1
1573	2223	9.52	410	1.2	23	80	210	5.8	2	2
1197	2219	9.55	240	0.0	23	86	139	7.1	2	1
1682	2161	8.18	480	1.7	26	135	365	0.9	1	2
1562	2137	9.42	450	0.0	33	81	210	5.9	2	2
1878	2122	9.60	550	0.0	16	49	119	4.5	2	2
895	2070	7.36	10	8.3	77	158	761	14.5	1	1
1044	2030	9.69	150	0.0	23	75	203	6.8	2	1
1429	2020	9.57	530	0.0	22	72	198	4.9	2	2
1670	2007	9.42	660	1.7	24	112	282	4.5	1	2
1747	2000	9.73	650	0.0	17	50	152	5.4	2	2
1867	1994	10.08	680	0.0	14	49	133	4.4	3	2
889	1933	7.20	20	7.0	65	141	659	14.8	1	1
1599	1891	8.86	690	2.1	32	162	375	5.5	1	2
924	1865	8.46	70	0.0	21	84	232	6.6	2	1
1535	1852	8.29	710	1.7	24	94	222	3.4	1	2
1719	1840	9.52	760	0.8	18	57	176	5.0	3	2
814	1794	7.22	10	6.8	66	144	643	13.3	1	1
1851	1773	9.09	1000	0.0	18	50	117	5.3	2	3
1117	1741	9.20	310	0.0	21	56	166	4.1	2	2
959	1723	7.82	160	2.2	27	131	317	4.5	2	1
846	1722	8.47	70	0.0	24	65	191	6.0	2	1
1562	1687	9.09	750	0.0	23	51	136	4.3	2	2
734	1666	7.36	20	7.4	72	181	801	15.2	1	1
1317	1625	9.97	540	0.8	46	42	141	4.5	3	2
874	1604	7.93	160	1.8	25	81	241	2.9	2	2
655	1564	5.18	20	6.6	75	173	784	11.4	1	1
1131	1545	9.63	390	0.0	24	48	128	7.1	2	3
1222	1542	10.13	480	1.0	29	48	158	5.2	3	2
1467	1485	9.87	760	0.4	23	48	143	6.6	2	3
611	1475	5.70	80	6.3	63	159	765	12.8	1	1
1601	1460	9.03	860	0.8	18	37	126	4.6	2	3
1728	1458	9.71	860	0.4	20	39	113	4.1	2	3
1057	1450	8.47	410	0.0	23	49	140	6.1	2	3
870	1425	5.80	270	7.8	75	399	1060	9.0	1	1
1391	1369	10.32	720	0.8	19	41	129	4.6	3	3
476	1305	6.34	20	10.8	85	333	1161	9.6	1	1
1676	1263	9.92	680	0.0	22	48	130	6.1	2	3
540	1255	5.76	80	3.9	47	268	703	7.0	1	1
1092	1233	7.64	560	0.7	22	45	119	3.6	1	1
381	1224	6.28	70	9.4	88	462	1383	8.5	1	1
1174	1221	8.84	630	0.0	30	67	221	5.7	2	3
1281	1212	10.52	650	0.4	25	84	240	8.8	2	3
422	1173	8.57	140	2.8	36	216	545	10.7	2	1
528	1167	6.48	130	3.5	46	252	676	6.2	1	1
465	1164	6.48	110	4.7	55	315	793	6.5	1	1
1514	1103	9.40	500	0.0	27	64	192	7.5	2	3

1252	1101	6.32	750	3.4	55	325	778	6.9	1	[illegible]
693	1097	8.17	320	0.4	22	76	186	6.5	2	1
312	1079	6.42	100	5.6	68	429	1136	8.2	1	1
1380	1073	8.76	500	0.4	23	63	203	7.2	2	2
862	1066	7.95	460	0.8	25	87	226	5.6	2	1
510	1058	6.32	260	3.1	39	237	593	7.0	1	1
353	1042	8.53	150	2.4	41	145	505	9.4	2	1
407	1027	6.32	200	3.9	49	260	685	5.7	1	1
580	1010	8.46	320	0.4	24	81	198	6.6	2	1
210	966	7.54	80	2.5	36	204	560	4.4	1	1
432	945	6.16	270	2.9	45	228	549	7.3	1	1
495	936	6.56	320	3.9	48	241	680	8.2	1	1
520	878	9.04	380	1.2	31	73	206	6.9	3	1
805	867	7.65	630	0.4	26	73	180	7.0	2	1
434	861	7.86	310	2.0	27	146	451	7.0	3	1
101	857	7.68	70	2.7	34	226	577	10.2	2	1
1182	840	7.79	380	0.4	22	49	157	6.4	2	1
275	816	8.50	220	2.6	33	163	420	9.0	2	1
458	810	6.90	360	2.7	36	201	539	4.3	1	1
1018	758	7.61	420	0.8	23	66	199	6.5	2	1
485	733	8.74	430	1.5	29	95	296	5.4	3	1
5	706	8.53	70	2.7	37	214	553	9.4	2	1
606	698	8.41	540	0.4	18	68	187	5.9	2	1
429	694	7.63	450	2.0	38	148	400	6.5	2	1
866	681	9.40	460	0.8	21	51	162	5.7	3	1
1721	666	8.70	80	3.7	53	250	722	9.1	2	2
1551	653	7.30	50	18.1	76	464	1672	17.0	1	1
203	649	8.65	280	1.8	27	129	332	7.0	2	1

References

AALDERS, H. J. G. L. (1996). Quality metrics for GIS. In M. J. Kraak and M. Molenaar (eds.), *Advances in GIS Research II*, Proc. 7th International Symposium on Spatial Data Handling, 12–16 Aug. 1996, Delft, The Netherlands, pp. 5B1–5B10.

ABEL, D. J. (1983). Towards a relational database for geographic information systems. Workshop on databases in the Natural Sciences. CSIRO Division of Computing Research. 7–9 Sept. 1983. Cunningham Laboratory, Brisbane, Queensland, Australia, 225 pp. mimeo.

—— and MARK, D. M. (1990). A comparative analysis of some two-dimensional orderings. *International Journal of Geographical Information Systems*, 4: 21–31.

ACRES, B. D., BOWER, R. P., BURROUGH, P. A., FOLLAND, C. J., KALSI, M. S., THOMAS, P., and WRIGHT, P. S. (1976). *The Soils of Sabah*. Land Resources Division, Ministry of Overseas Development, London, 5 vols.

ADAMS, J., PATTON, C., READER, C., and ZAMORA, D. (1984). Fast hardware for geometric warping. *Proc. 3rd Australian Remote Sensing Conference*, Queensland.

AGI (1996). *AGI Online Glossary of GIS*. The Internet. (*http://www.geo.ed.ac.uk/root/agidict/html/welcome.html*)

ALDENDERFER, M., and MASCHNER, H. D. G. (1996). *Anthropology, Space and Geographic Information Systems*. Oxford University Press, New York, 294 pp.

ALDRED, B. K. (1972). Point in polygon algorithms. Peterlee UK, IBM Ltd.

ALMELIO, C. F. (1974). Charge coupled devices. *Scientific American*, 230 (2): 22–31.

ALONSO, W. (1968). Predicting best with imperfect data. *J. Am. Inst. Planners* (APA Journal), July: 248–55.

ALTMAN, D. (1994). Fuzzy set theoretic approaches for handling imprecision in spatial analysis. *International Journal of Geographical Information Systems*, 8: 271–90.

American Society of Photogrammetry (1960). *Manual of Aerial Photo Interpretation*, Washington, DC.

American Society of Photogrammetry and Remote Sensing (1980). *Manual of Photogrammetry*, Washington, DC.

ARCTUR, D., and WOODSFORD, P. (1996). *Introduction to Object-Oriented GIS Technology Workshop Outline*. Laser-Scan, Cambridge.

ARMSTRONG, M., and MATHERON, G. (1986*a*). Disjunctive kriging revisited: Part I. *Mathematical Geology*, 18: 711–27.

—— —— (1986*b*). Disjunctive kriging revisited: Part II. *Mathematical Geology*, 18: 729–41.

ARMSTRONG, M. P. and HOPKINS, L. D. (1983). Fractal enhancement for thematic display of topologically stored data. Proc. AUTOCARTO 6, Ottawa, Canada.

ARONOFF, S. (1989). *Geographic Information Systems: A Management Perspective.* WDL Publications, Ottawa, Canada.

AYENI, O. O. (1982). Optimum sampling for Digital Terrain Models. *Photogrammetric Engineering and Remote Sensing*, 48(11): 1687–94.

BAND, L. E. (1986). Topographic partition of watersheds with digital elevation models. *Water Resources Research*, 22: 15–24.

BARROW, J. D. (1992). *Pi in the Sky*. Oxford University Press, Oxford, 317 pp.

BATTY, M., and XIE, Y. (1994*a*). Modelling inside GIS: Part 1. Model structures, exploratory spatial data analysis, and aggregation. *International Journal of Geographical Information Systems*, 8: 291–307.

—— —— (1994*b*). Modelling inside GIS: Part 2. Selecting and calibrating urban models using ARC-INFO. *International Journal of Geographical Information Systems*, 8: 451–70.

BECKETT, P. H. T. (1971). The cost-effectiveness of soil survey. *Outlook in Agriculture*, 6(5): 191–8.

—— and BURROUGH, P. A. (1971). The relation between cost and utility in soil survey IV. *Journal of Soil Science*, 22: 466–80.

—— and WEBSTER, R. (1971). Soil variability—a review. *Soils and Fertilisers*, 34: 1–15.

BEEK, K. J. (1978). *Land Evaluation for Agricultural Development*. International Institute for Land Reclamation and Improvement, Pub. 23, Wageningen.

BEERS, B. J. (1995). *FRANK: The Design of a New Landsurveying System Using Panoramic Images*. Delft University Press, Delft.

BERGE, H. F. M. TEN, STROOSNIJDER, L., BURROUGH, P. A., BREGT, A. K., and DE HEUS, M. J. (1983). Spatial variability of physical soil properties influencing the temperature of the soil surface. *Agricultural Water Management*, 6: 213–26.

BERRY, J. K. (1993). *Beyond Mapping: Concepts, Algorithms, and Issues in GIS*. GIS World Books, GIS World, Inc. Fort Collins, Color. 246 pp.

BEVEN, K., and KIRKBY, M. J. (1979). A physically-based, variable contributing area model of basin hydrology. *Hydrological Science Bulletin*, 24: 1–10.

—— and Moore, I. D. (eds.) (1994). *Terrain Analysis and Distributed Modelling in Hydrology*. Advances in Hydrological Processes, Wiley, Chichester, 249 pp.

—— —— Scholfield, N., and Tagg, A. F. (1984). Testing a physically-based flood forecasting model (TOPMODEL) for three UK catchments. *Journal of Hydrology*, 69: 119–43.

Beyer, R. I. (1984). A database for a botanical plant collection. In B. M. Evans (ed.), *Computer-aided Landscape Design: Principles and Practice*. The Landscape Institute, Scotland, pp. 134–41.

Bezdek, C. J. (1981). *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum Press, New York, 256 pp.

—— Erhlich, R., and Full, W. (1984). FCM: the fuzzy c-means clustering algorithm. *Computers and Geosciences*, 10: 191–203.

Bhalla, N. (1991). Object-oriented data models: a perspective and comparative review. *Journal of Information Science*, 17: 145–60.

Bie, S. W., and Beckett, P. H. T. (1970). The costs of soil survey. *Soils and Fertilisers*, 33: 203–17.

—— —— (1973). Comparison of four independent soil surveys by air-photo interpretation, Paphos area (Cyprus). *Photogrammetria*, 29: 189–202.

Bierkens, M. F. P. (1994). *Complex Confining Layers: A Stochastic Analysis of Hydraulic Properties at Various Scales*. Royal Dutch Geographical Association / Faculty of Geographical Sciences, Utrecht University, Utrecht.

—— and Burrough, P. A. (1993*a*). The indicator approach to categorical soil data. (I) Theory. *Journal of Soil Science*, 44: 361–68.

—— —— (1993*b*). The indicator approach to categorical soil data. (II) Application to mapping and landuse suitability analysis. *Journal of Soil Science*, 44: 369–81.

Bijker, W. E., Hughes, T. P., and Pinch, T. J. (eds.) (1987). *The Social Construction of Technological Systems: New Directions in the Sociology and History of Technology*. MIT Press, Cambridge, Mass.

Blakemore, M. (1984). Generalisation and error in spatial data bases. *Cartographica*, 21: 131–9.

Boerma, P. N., Henneman, G. R., Kauffman, J. H., and Verwey, H. E. (1974). Detailed soil survey of the Marongo area. Preliminary Report No. 3. Training project in Pedology. Agricultural University, Wageningen.

Bolstad, P. V., Gessler, P., and Lillesand, T. M. (1990). Positional uncertainty in manually digitized map data. *International Journal of Geographical Information Systems*, 4: 399–411.

Bornand, M., Legros, J. P., and Moinereau, J. (1977). *Carte Pédologique de France. N.19: Privas*. Versailles, Centre Nationale de Recherche Agronomiques (CNRA).

Bouillé, F. (1978). Structuring cartographic data and spatial processes with the hypergraph-based data structure. In G. Dutton (ed.), *First International Advanced Study Symposium on Topological Data Structures for Geographic Information Systems*, vol. v. Laboratory for Computer Graphics and Spatial Analysis, Harvard, Cambridge, Mass., 17–21 Oct. 1977.

Bouma, J., and Bell, J. P. (eds.) (1983). *Spatial Variability*. Special Issue, *Agricultural Water Management*, 6 (2/3).

—— and Bregt, A. K. (1989). *Land Qualities in Space and Time*, Proc. Symposium organized by the International Society of Soil Science (ISSS), Wageningen, The Netherlands, 22–6 Aug. 1988. PUDOC, Wageningen, 356 pp.

Boyle, A. R. (1981). Concerns about the present applications of computer-assisted cartography. *Cartographica*, 18: 31–3.

—— (1982). The last ten years of automated cartography: a personal view. In D. Rhind and T. Adams (eds.), *Computers in Cartography*. British Cartographic Society, London, pp. 1–3.

Brandenburger, A. J., and Gosh, S. K. (1985). The world's topographic and cadastral mapping operation. *Photogrammetric Engineering and Remote Sensing*, 51: 437–44.

Bregt, A. K., and Beemster, J. G. R. (1989). Accuracy in predicting moisture deficits and changes in yield from soil maps. *Geoderma*, 43: 301–10.

—— Denneboom, J., Gesink, H. J., and van Randen, Y. (1991). Determination of rasterizing error: a case study with the soil map of the Netherlands. *International Journal of Geographical Information Systems*, 5: 361–8.

Brinkman, R., and Smyth, A. J. (1973). *Land Evaluation for Rural Purposes*. Summary of an Expert Consultation, Wageningen, the Netherlands, 6–12 Oct. 1972. International Institute for Land Reclamation and Improvement, Wageningen, Pub. 17.

Brunt, M. (1967). The methods employed by the Directorate of Overseas Surveys in the Assessment of Land Resources. *Actes du 2me Symp. Intil. de Photo Interpretation*, Paris 1966.

Buiten, H. J., and Clevers, J. G. P. W. (1993). *Land Observation by Remote Sensing: Theory and Applications*. Gordon & Breach, Reading, UK.

Bullock, A., and Stallybrass, O. (1977). *Fontana Dictionary of Modern Thought*. Fontana Books, London.

Burgess, T. M., and Webster, R. (1980). Optimal interpolation and isarithmic mapping of soil properties I. The semivariogram and punctual kriging. *Journal of Soil Science*, 31: 315–31.

—— —— (1984). Optimal sampling strategies for mapping soil types. I. Distribution of boundary spacings. *Journal of Soil Science*, 35: 641–54.

Burrough, P. A. (1969). Studies in soil survey methodology. Unpublished D.Phil. Thesis, Oxford University.

References

BURROUGH, P. A. (1975). Message sticks used by Murut and Dusun people in Sabah. *Journal of the Malaysian Branch of the Royal Asiatic Society*, 48(2): 119–23.

—— (1980). The development of a landscape information system in the Netherlands, based on a turn-key graphics system. *Geoprocessing*, 1: 257–74.

—— (1983*a*). Multiscale sources of spatial variation in soil I. The application of fractal concepts to nested levels of soil variation. *Journal of Soil Science*, 34: 577–97.

—— (1983*b*). Multiscale sources of spatial variation in soil II. A non-Brownian fractal model and its application in soil survey. *Journal of Soil Science*, 34: 599–620.

—— (1984). The application of fractal ideas to geophysical phenomena. *Bull. Inst. Mathematics and Its Applications*, 20(3/4): 36–42.

—— (1985). Fakes, facsimiles and fractals: fractal models of geophysical phenomena. In S. Nash (ed.), *Science and Uncertainty*, IBM UK Ltd/Science Reviews, Northwood, pp. 151–70.

—— (1986). *Principles of Geographical Information Systems for Land Resources Assessment*. Oxford University Press, Oxford, 194 pp.

—— (1989). Fuzzy mathematical methods for soil survey and land evaluation. *Journal of Soil Science*, 40: 477–92.

—— (1991*a*). Sampling designs for quantifying map unit composition. Chapter 7 in M. Mausbach and L. Wilding (eds), *Spatial Variabilities of Soils and Landforms*, Soil Science Society of America Special Publication No. 28, Madison, Wis. pp. 89–125.

—— (1991*b*). Soil information systems. Chapter IIc.1 in D. J. Maguire, M. F. Goodchild, and D. W. Rhind (eds.), *Geographical Information Systems: Principles and Applications*, vol. ii. Longman Scientific and Technical, Harlow, pp. 153–69.

—— (1992). Development of intelligent geographical information systems. *International Journal of Geographical Information Systems*, 6: 1–12.

—— (1993*a*). Fractals and Geostatistical Methods in Landscape Studies. In N. Lam and Lee de Cola (eds.), *Fractals in Geography*, Prentice Hall, Englewood Clifts, NJ, pp. 87–121.

—— (1993*b*). Soil variability: a late 20th century view. *Soils & Fertilizers*, 56: 529–62.

—— (1996*a*). Opportunities and limitations of GIS-based modeling of solute transport at the regional scale. Chapter 2 in D. L. Corwin and K. Loague (eds.), Special SSSA Publication *Application of GIS to the Modeling of Non-Point Source Pollutants in the Vadose Zone*. American Society of Agronomy.

—— (1996*b*). A European view of Global Spatial Data Infrastructure. *Proc. Emerging Global Spatial Data Infrastructure*. ed. C. Chenez. A Conference held under the patronage of Dr Martin Bangemann, European Commission, EUROGI/Deutscher Dachverband für Geoinformation/Atlantic Institute/Open GIS Consortium/ Federal Data Committee/Fédération Internationale des Géomètres (Commission 3), 4–6 Sep. 1996, Königswinter, Germany.

—— and FRANK, A. U., (eds.) (1996). *Geographical Objects with Indeterminate Boundaries*. Taylor & Francis, London, 345 pp.

—— and MASSER, F. I. (1998). *European Geographic Information Infrastructures: Opportunities and Pitfalls*. Taylor & Francis, London.

—— and WEBSTER, R. (1976). Improving a reconnaissance soil classification by multivariate methods. *Journal of Soil Science*, 27: 554–71.

—— BECKETT, P. H. T., and JARVIS, M. G. (1971). The relation between cost and utility in soil survey I–III. *Journal of Soil Science*, 22: 359–94.

—— BROWN, L., and MORRIS, E. C. (1977). Variations in vegetation and soil pattern across the Hawkesbury Sandstone plateau from Barren Grounds to Fitzroy Falls, New South Wales. *Australian J. Ecol.* 2: 137–59.

—— GILLESPIE, M., HOWARD, B., and PRISTER, B. (1996). Redistribution of ^{137}Cs in Ukraine wetlands by flooding and runoff. In K. Kovar and H. P. Nachnebel, *Application of Geographic Information Systems in Hydrology and Water Resources Management*. IAHS Publications No. 235, IAHS Press, Institute of Hydrology, Wallingford, pp. 269–78.

—— MACMILLAN, R. A., and VAN DEURSEN, W. P. A. (1992). Fuzzy classification methods for determining site suitability from soil profile observations and topography. *Journal of Soil Science*, 43: 193–210.

—— VAN GAANS, P., and HOOTSMANS, R. J. (1997). Continuous classification in soil survey: spatial correlation, confusion and boundaries. *Geoderma*, 77: 115–136.

CARRARA, A., BITELLI, G., and CARLA, R. (1997). Comparison of techniques for generating digital terrain models from contour lines. *International Journal of Geographical Information Science*, 11: 451–74.

CARTER, J. R. (1989). On defining the geographic information system. In W. J. Ripple (ed.), *Fundamentals of Geographical Information Systems: A Compendium*. ASPRS/ ACSM, Falls Church, Va., pp. 3–7.

CARVER, S. J. (1991). Integrating multi-criteria evaluation with geographical information systems. *International Journal of Geographical Information Systems*, 5: 321–40.

—— and BRUNSDON, C. F. (1994). Vector to raster conversion error and feature complexity: an empirical study using simulated data. *International Journal of Geographical Information Systems*, 8: 261–70.

CERUTI, A. (1980). A method of drawing slope maps from contour maps by automatic data acquisition and processing. *Computers and Geosciences*, 6: 289–97.

CHANCE, A., NEWELL, R. G., and THERIAULT, D. G. (1995). Smallworld GIS: an object-oriented GIS—issues and solutions. Smallworld Technical Paper, paper no. 3.

CHRISMAN, N. R. (1984*a*). The role of quality information in the long-term functioning of a geographic information system. *Cartographica* 21: 79–87.

—— (1984*b*). On storage of coordinates in geographic information systems. *Geo-Processing*, 2: 259–70.

CHRISTIAN, C. S., and STEWART, G. A. (1968). Aerial surveys and integrated studies. *Proc. Toulouse Conference, 21–8 Sept. 1964. Natural Resources Research VI*, UNESCO, Paris.

CLIFF, A. D., and ORD, J. K. (1981). *Spatial processes: Models and applications.* Pion, London.

CMD (1984). *Annales météorologiques Ardèche 1984.* Comité Météorologiques Départmentate, Aubenas, France.

COOK, B. G. (1978). The structural and algorithmic basis of a geographic database. *Harvard Papers on Geographic Information Systems: First International Advanced Study Symposium on Topological Data Structures for Geographic Information Systems*, vol. iv, ed. G. Dutton. Laboratory for Computer Graphics and Spatial Analysis, Graduate School of Design, Harvard University.

—— (1983). An introduction to the design of geographic databases. In *Proc. Workshop on databases in the Natural Sciences*, CSIRO Division of Computing Research, 7–9 Sept. 1983, Cunningham Laboratory, Brisbane, Queensland, 175–86.

CORINE (1992). *Corine Land Cover Project.* Environmental Agency, Copenhagen.

COSTA-CABRAL, M., and BURGES, S. J. (1993). Digital Elevation Model Networks (DEMON): A model of flow over hillslopes for computation of contributing and dispersal areas. *Water Resources Research*, 30(6): 1681–92.

COUCLELIS, H. (1992). People manipulate objects (but cultivate fields): beyond the raster-vector debate in GIS. In A. U. Frank, I. Campari, and Formentini, U. (eds.), *Theories and Methods of Spatio-Temporal Reasoning in Geographic Space.* Lecture Notes in Computer Science 639. Springer Verlag, Berlin, pp. 65–77.

COWEN, D. J. (1988). GIS versus CAD versus DBMS: what are the differences? *Photogrammetric Engineering and Remote Sensing*, 54: 1551–4.

CRESSIE, N. A. C. (1991). *Statistics for Spatial Data.* Wiley, New York, 900 pp.

—— and HAWKINS, D. M. (1980). Robust estimation of the variogram. *Mathematical Geology*, 12: 115–25.

DAHL, O.-J., and NYGAARD, K. (1966). SIMULA—an algol-based simulation language. *Comm ACM*, 9: 671–78.

DALE, P. F., and MCLAUGHLIN, J. D. (1988). *Land Information Management.* Oxford University Press, Oxford, 266 pp.

DATE, C. J. (1995). *An Introduction to Database Systems*, 6th edn., Addison-Wesley, Reading, Mass.

DAVID, B., VAN DEN HERREWEGEN, M., and SALGÉ, F. (1996). Conceptual models for geometry and quality of geographic information. In P. A. Burrough and A. U. Frank (eds.). *Geographic Objects with Indeterminate Boundaries.* Taylor & Francis, London.

DAVIDSON, D. A., THEOCHAROPOULOS, S. P., and BLOKSMA, R. J. (1994). A land evaluation project in Greece using GIS and based on Boolean and fuzzy set methodologies. *International Journal of Geographical Information Systems*, 8: 369–84.

DAVIS, J. C. (1986). *Statistics and Data Analysis in Geology*, 2nd edn. Wiley, New York, 646 pp.

DAVIS, L. S. (1976). A survey of edge-detection techniques. *CGIP*, 4 (3): 248–70.

DE BAKKER, H. (1979). *Major Soils and Soil Regions in The Netherlands.* Junk, The Hague/PUDOC, Wageningen, 203 pp.

DE FLORIANI, L., and MAGILLO, P. (1994). Visibility algorithms on triangulated digital terrain models. *International Journal of Geographical Information Systems*, 8: 13–41.

DE, GRUIJTER, J. J., and MCBRATNEY, A. B. (1988). A modified fuzzy *k*-means method for predictive classification. In H. H. Bock (ed.), *Classification and Related Methods of Data Analysis.* Elsevier, Amsterdam, pp. 97–104.

—— and MARSMAN, B. (1985). Transect sampling for reliable information on mapping units. In J. Bouma and D. Nielsen (eds.), *Spatial Analysis of Soil Data.* PUDOC, Wageningen.

—— and TER BRAAK, C. J. F. (1990). Model-free estimation from spatial samples: a reappraisal of Classical Sampling Theory. *Mathematical Geology*, 22: 407–15.

—— WALVOORT, D., and VAN GAANS, P. F. M. (1997). Continuous soil maps—a fuzzy set approach to bridge the gap between aggregation levels of process and distribution models. *Geoderma*, 77: 169–96.

DE JONG, S. M. (1994). *Applications of Reflective Remote Sensing for Land Degradation Studies in a Mediterranean Environment.* Netherlands Geographical Studies 177, Koninklijk Nederlands Aardrijkskundig Genootschap, University of Utrecht, 237 pp.

—— and RIEZEBOS, H. TH. (1997). SEMMED: a distributed approach to soil erosion modelling. In A. Spiteri (ed.), *Remote Sensing 96. Integrated Applications for Risk Assessment and Disaster Prevention for the Mediterranean.* Balkema, Rotterdam.

DELL'ORCO, P., and GHIRON, M. (1983). Shape representation by rectangles preserving fractality. Proc: AUTOCARTO, 6 Oct. 1983, Ottawa.

Department of Environment (DoE) (1987). *Handling Geographic Information.* HMSO, London.

References

De Roo, A. P. J., Hazelhoff, L., and Burrough, P. A. (1989). Soil erosion modelling using 'ANSWERS' and geographical information systems. *Earth Surface Processes and Land Forms*, 14: 517–32.

—— —— and Heuvelink, G. B. M. (1992). Estimating the effects of spatial variability of infiltration on the output of a distributed runoff and soil erosion model using Monte Carlo methods. *Hydrological Processes*, 6: 127–43.

Desmet, P. J. J. (1997). Effects of Interpolation errors on the analysis of DEMs. *Earth Surface Processes and Landforms*, 22, 563–580.

—— and Govers, G. (1996). Comparison of routing algorithms for digital elevation models and their implications for predicting ephemeral gullies. *International Journal of Geographical Information Systems*, 10: 311–31.

De Soto, H. (1993). The missing ingredient, *Economist*, 11 Sept.

Deutsch, C., and Journel, A. G. (1992). GSLIB Geostatistical Handbook. Oxford University Press, New York.

Dilke, O. A. W. (1971). *The Roman Land Surveyors. An Introduction to the Agrimensores.* David and Charles, Newton Abbot.

Douglas, D. H., and Peuker, T. K. (1973). Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. *Canadian Cartographer*, 10: 112–22.

Dozier, J. (1980). A clear-sky spectral solar radiation model for snow-covered mountainous terrain. *Water Resources Research*, 16: 709–18.

—— and Frew, J. (1990). Rapid calculation of terrain parameters for radiation modelling from digital elevation data. *IEEE Transactions on Geoscience and Remote Sensing*, 28: 963–9.

Dubayah, R. (1992). Estimating net solar radiation using Landsat Thematic Mapper and digital elevation data. *Water Resources Research*, 28: 2469–84.

—— and Rich, P. M. (1995). Topographic solar radiation models for GIS. *International Journal of Geographical Information Systems*, 9: 405–19.

Dubrule, O. (1983). Two methods with different objectives: splines kriging. *Mathematical Geology*, 15: 245–55.

—— (1984). Comparing splines & kriging. *Computers & Geosciences*, 10: 327–38.

Duffield, B. S., and Coppock, J. T. (1975). The delineation of recreational landscapes: the role of a computer-based information system. *Trans. Inst. British Geographers*, 66: 141–8.

Dunn, R., Harrison, A. R., and White, J. C. (1990). Positional accuracy and measurement error in digital databases of land use: an empirical study. *International Journal of Geographical Information Systems*, 4: 385–97.

Dutton, G. H. (1981). Fractal enhancement of cartographic line detail. *American Cartographer*, 8(1): 23–40.

—— (1996). Encoding and handling geospatial data with hierarchical triangular meshes. In M. J. Kraak and M. Molenaar (eds.), *Advances in GIS Research II. Proceedings 7th International Symposium on spatial Data Handling*, 12–16 Aug. 1996, Delft, The Netherlands, pp. 8B15–28.

Egenhofer, M. J., and Herring, J. R. (eds.) (1995). *Advances in Spatial Databases*. Springer Verlag, Berlin.

Elwell, H. A., and Stocking, M. A. (1982). Developing a simple yet practical method of soil-loss estimation. *Tropical Agriculture*, 59: 43–8.

Englund, E. J. (1990). A variance of geostatisticians. *Mathematical Geology*, 22: 417–55.

Estes, J. (1995). What GIS is, where it came from and what it does. *Proc. Cambridge Conference for National Mapping Organisations 1995*, Workshop Paper 2, 1/18.

—— Hajic, E., and Tinney, L. R. (1983). Fundamentals of image analysis: analysis of visible and thermal infrared data. Chapter 24 of R. N. Colwell (ed.), *Manual of Remote Sensing*. American Society of Photogrammetry, Falls Church, Va., 2440 pp.

EUROSTAT (1996). The spatial dimension of the European statistical system. EUROSTAT, Luxembourg.

Evans, I. S. (1977). The selection of class intervals. *Transactions Institute of British Geographers* (ns), 2: 98–124.

—— (1980). An integrated system of terrain analysis and slope mapping. *Zeitschrift für Geomorphologie Suppl.* 36: 274–95.

Fabos, J. G., and Caswell, S. J. (1977). Composite landscape assessment: metropolitan landscape planning model METLAND. *Res. Bull.* 637. Massachusetts Agric. Experimental Station, University of Massachusetts at Amherst.

FAO (1976). A framework for land evaluation. *Soils Bull.* 32. FAO Rome and Int. Inst. Land Reclam. Improv. Pub. 22.

Fedra, K. (1996). Distributed models and embedded GIS: strategies and case studies of integration. In M. F. Goodchild, L. T. Steyaert, B. O. Parks, C. Johnston, D. Maidment, M. Crane, and S. Glendinning (eds.), *GIS and Environmental Modeling: Progress and Research Issues.* GIS World Books, Fort Collins, Colo., pp. 413–18.

Fisher, H. T. (1978). Thematic cartography: what it is and what is different about it. *Harvard Papers in Theoretical Cartography*. Laboratory for Computer Graphics and Spatial Analysis, Harvard, Cambridge, Mass.

Fisher, P. F. (1991). Modelling soil map-unit inclusions by Monte Carlo simulation. *International Journal of Geographical Information Systems*, 5: 193–208.

—— (1995). An exploration of probable viewsheds in landscape planning. *Environment and Planning B: Planning and Design*, 22: 527–46.

—— (1996). Reconsideration of the viewshed function in terrain modelling. *Geographical Systems*, 3: 33–58.

Fitzgerald, R. W., and Lees, B. G. (1996). Temporal context in floristic classification. *Computers and Geosciences*, 22: 981–94.

Fix, R. A., and Burt, T. P. (1995). Global Positioning System: an effective way to map a small area or catchment. *Earth Surface Processes and Landforms*, 20: 817–27.

Fleming, M. D., and Hoffer, R. M. (1979). Machine Processing of Landsat MSS Data and DMA topographic data for forest cover type mapping *LARS Technical Report 062879*, Laboratory for Applications of Remote Sensing, Purdue University, West Lapayette, Indiana.

Frank, A. U. (1988). Requirements for a database management system for a GIS. *Photogrammetric Engineering and Remote Sensing*, 54: 1557–64.

—— and Campari, I. (eds.) (1993). *Spatial Information Theory: A Theoretical Basis for GIS*. Springer Verlag, Berlin.

—— —— and Formentini, U. (eds.) (1992). *Theories and Methods of Spatio-Temporal Reasoning in Geographic Space*. Lecture Notes in Computer Science 639. Springer Verlag, Berlin, pp. 65–77.

Franklin, W. R. (1984). Cartographic errors symptomatic of underlying algebra problems. Proc. Int. Symposium on Spatial Data Handling, 20–4 Aug. 1984, Zurich, Switzerland, pp. 190–208.

Frapporti, G., Vriend, S. P., and van Gaans, P. F. M. (1993). Hydrogeochemistry of the shallow Dutch ground water: interpretation of the national ground water quality monitoring network. *Water Resources Research*, 29: 2993–3004.

Freeman, G. T. (1991). Calculating catchment area with divergent flow based on a regular grid. *Computers and Geosciences*, 17: 413–22.

Freeman, H. (1974). Computer processing of line-drawing images. *Computing Surveys*, 6: 54–97.

Frolov, Y. S., and Maling, D. H. (1969). The accuracy of area measurements by point counting techniques. *Cartographic J.* 6: 21–35.

Gaans, P. F. M. van, Vriend, S. P., van der Wal, J., and Schuiling, R. D. (1986). Integral rock analysis, a new approach to lithogeochemical exploration. Application: Carboniferous sediments of a coal exploration drilling, Limburg, the Netherlands. Report to the Commission of the European Communities, contract MSM-073-NL(N), 87 pp.

—— —— and Frapporti, G. (1992). Oxidation of pyrite in the subsoil of Noord-Brabant. Setting, causes, and effects for ground water quality. *H2O*, 25: 736–45 (in Dutch).

Gahegan, M. N., and Roberts, S. A. (1988). An intelligent, object-oriented geographical information system. *International Journal of Geographical Information Systems*, 2(2): 101–10.

Gee, D. M., Anderson, A. G., and Baird, L. (1990). Large-scale floodplain modelling. *Earth Surface Processes and Landforms*, 15: 513–23.

Geertman, S. C. M., and Ritsema Van Eck, J. R. (1995). GIS and models of accessibility potential: an application in planning. *International Journal of Geographical Information Systems*, 9: 67–80.

Goldberg, A., and Robson, D. (1983). *Smalltalk-80*. Addison-Wesley, Reading, Mass.

Goméz-Hernández, J. J., and Journel, A. G. (1992). Joint sequential simulation of multigaussian fields. In A. Soares (ed.), *Proc. Fourth Geostatistics Congress, Troia, Portugal*. Quantitative Geology and Geostatistics (5), Kluwer Academic Publishers, Dordrecht, pp. 85–94.

—— and Srivastava, R. M. (1990). ISIM3D: an ANSI-C three-dimensional multiple indicator conditional simulation program. *Computers and Geosciences*, 16(4): 395–440.

Goodchild, M. F. (1978), Statistical aspects of the polygon overlay problem. In G. Dutton (ed.), *Harvard Papers on Geographic Information Systems* vi. Addison-Wesley, Reading, Mass.

—— M. F. (1980). Fractals and the accuracy of geographical measures. *Mathematical Geology*, 12: 85–98.

—— (1989). Modeling error in objects and fields. In M. F. Goodchild and S. Gopal (eds.), *Accuracy of Spatial Databases*. Taylor & Francis, London, pp. 107–13.

—— and Gopal, S. (1989). *The Accuracy of Spatial Databases*. Taylor & Francis, London.

—— Parks, B. O., and Steyaert, L. T. (1993). *Environmental Modeling with GIS*. Oxford University Press, Oxford.

—— Steyaert, L. T., Parks, B. O., Johnston, C., Maidment, D., Crane, M., and Glendinning, S. (eds.) (1996). *GIS and Environmental Modeling: Progress and Research Issues*. GIS World Books, Fort Collins, Colo., 486 pp.

Goode, J. (ed.) (1997). Precision agriculture: spatial and temporal variability of environmental quality. Wiley, Chichester (Ciba Foundation Symposium 210).

Goudie, A. S., Atkinson, B. W., Gregory, K. J., Simmons, I. G., Stoddart, D. R., and Sugden, D. (eds.) (1988). *The Encyclopaedic Dictionary of Physical Geography*. Blackwell Reference, Oxford, 528 pp.

Graham, I. (1994). *Object Oriented Methods*. 2nd edn., Addison-Wesley, Wokingham.

Grothenn, D., and Stanfenbiel, W. (1976). Das Standarddatenformat zum Austausch Kartografischer Daten. *Nachrichten aus dem Karten- und Vermessungswesen*, 1(69): 25–49.

Gruenberger, F. (1984). Computer recreations. *Scientific American*, 250(4): 10–14.

Gunnink, J. L., and Burrough, P. A. (1997). Interactive spatial analysis of soil attribute patterns using Explanatory Data Analysis (EDA) and GIS. In M. Fischer, H. J.

Scholten, and D. Unwin (eds.), *Spatial Analytical Perspectives on GIS*. Taylor & Francis, London, pp. 87–100.

Guptill, S. C. (1991). Spatial data exchange and standardization. In D. J. Maguire, M. F. Goodchild, and D. W. Rhind (eds.), *Geographical Information Systems, i: Principles*. Longman Scientific and Technical, Harlow, Essex, pp. 515–30.

—— and Morrison, J. (1995). (eds.). *The Elements of Spatial Data Quality*. Elsevier, Amsterdam.

Gutowitz, H. (1991). *Cellular Automata: Theory and Experiment*. MIT Press, Cambridge, Mass.

Guttman, A. (1984). R-trees: a dynamic index structure for spatial searching. *Proceedings of the 13th Association for Computing Machinery SIGMOD Conference*. ACM Press, Boston, New York.

Haining, R. (1990). *Spatial Data Analysis in the Social and Environmental Sciences*. Cambridge University Press, Cambridge.

Harvey, D. (1969). *Explanation in Geography*. Edward Arnold, London.

Haslett, J., Wills, G., and Unwin, A. (1990). SPIDER—an interactive statistical tool for the analysis of spatially distributed data. *International Journal of Geographical Information Systems*, 4: 285–96.

Hawkins, D. M., and Merriam, D. F. (1973). Optimal zonation of digitized sequential data. *Math. Geol.* 5: 389–95.

—— (1974). Zonation of multivariate sequences of digitized geologic data. *Math. Geol.* 6: 263–9.

Hearnshaw, H. M., and Unwin, D. J. (1994). *Visualization in Geographical Information Systems*. Wiley, Chichester, 259 pp.

Herring, J. R. (1992). TIGRIS: a data model for an object-oriented geographic information system. *Computers and Geosciences*, 18(4): 443–8.

Heuvelink, G. B. M. (1993). *Error Propagation in Quantitative Spatial Modelling*. KNAG, University of Utrecht Pub. 163, 160 pp.

—— and Burrough, P. A. (1993). Error propagation in cartographic modelling using Boolean logic and continuous classification. *International Journal of Geographical Information Systems*, 7: 231–46.

—— —— and Stein, A. (1989). Propagation of errors in spatial modelling with GIS. *International Journal of Geographical Information Systems*, 3: 303–22.

Hills, G. A. (1961). The ecological basis for land use planning. *Res. Rep. 46*, Ontario Department of Lands and Forests, Canada.

Hodgkiss, A. G. (1981). *Understanding maps*. Dawson, Folkstone.

Hodgson, M. E. (1995). What cell size does the computed slope/aspect angle represent? *Photogrammetric Engineering and Remote Sensing*, 61: 513–17.

Holroyd, F., and Bell, S. B. M. (1992). Raster GIS: Models of raster encoding. *Computers and Geosciences*, 18: 419–26.

Hopkins, L. D. (1977). Methods for generating land suitability maps: a comparative evaluation. *American Institute of Planners Journal* (Oct): 386–400.

Horn, B. K. P. (1981). Hill shading and the reflectance map. *Proc. IEEE* 69(1): 14–47.

Hutchinson, M. F. (1989). A new procedure for gridding elevation and stream line data with automatic removal of spurious pits. *Journal of Hydrology*, 106: 211–32.

—— (1995). Interpolating mean rainfall using thin plate smoothing splines. *International Journal of Geographical Information Systems*, 9: 385–404.

Huxhold, W. E. (1991). *An Introduction to Urban Geographic Information Systems*. Oxford University Press, New York, 337 pp.

—— and Levinsohn, A. G. (1995). *Managing Geographic Information System Projects*. Oxford University Press, New York, 247 pp.

IGN (1985). *Carte topographique 1 : 25.000*. Institute Géographique National, Paris.

Isaaks, E. H., and Srivastava, R. M. (1989). *An Introduction to Applied Geostatistics*. Oxford University Press, New York, 561 pp.

Jackson, M. J., and Woodsford, P. A. (1991). GIS data capture hardware and software. In D. J. Maguire, M. F. Goodchild, and D. W. Rhind (eds.), *Geographical Information Systems, i: Principles*. Longman Scientific and Technical, Harlow, pp. 239–49.

Jacobsen, I., Christerson, M., Jonsson, P., and Övergaard, G. (1992). *Object-Oriented Software Engineering: A Use Case Driven Approach*. Addison-Wesley, Wokingham, 524 pp.

Jenks, G. F. (1981). Lines, computers, and human frailties. *Annals AAG* 71(1): 1–10.

Jenson, S. J., and Domingue, J. O. (1988). Extracting topographic structure from digital elevation data for geographic information system analysis. *Photogrammetric Engineering and Remote Sensing*, 54: 1593–600.

Jetten, V. (1994). Modelling the effects of logging on the water balance of a tropical rain forest: a study in Guyana. *Tropenbos Series 6*, University of Utrecht, 183 pp.

Johnston, R. J., Gregory, D., and Smith, D. M. (eds.) (1988). *The Dictionary of Human Geography*. Blackwell Reference, Oxford.

Jones, K. H. (1997). A comparison of algorithms used to compute hill slopes and aspects as a property of the DEM. (pers. comm.).

Jongman, R. H. G., ter Braak, C. J. F., and van Tongeren, O. F. R. (1995). *Data Analysis in Community and Landscape Ecology*. Cambridge University Press, Cambridge.

JOURNEL, A. G. (1996). Modelling uncertainty and spatial dependence: stochastic imaging. *International Journal of Geographical Information Systems*, 10: 517–22.

—— and HUIJBREGTS, C. J. (1978). *Mining Geostatistics.* Academic Press, London.

—— and POSA, D. (1990). Characteristic behavior and order relations for indicator variograms. *Mathematical Geology*, 22: 1011–25.

KANDEL, A. (1986). Fuzzy mathematical techniques with applications. Addison-Wesley, Reading, Mass.

KARAOGLOU, A., DESMET, G., KELLY, G. N., and MENZEL, H. G. (eds.) (1996). The radiological consequences of the Chernobyl Accident. Proceedings of the First International Congress, Minsk, Belarus, EUR 16544 EN, Brussels-Luxembourg.

KAUFFMAN, A. (1975). *Introduction to the Theory of Fuzzy Subsets.* Academic Press, New York.

KELLY, R. E., MCCONNELL, P. R. H., and MILDENBERGER, S. J. (1977). The Gestalt photomapping system. *Photogrammetric Engineering & Remote Sensing*, 43: 1407–17.

KENNEDY, M. (1996). *The Global Positioning System and GIS.* Ann Arbor Press, Inc., Ann Arbor, 268 pp.

KIDNER, D. B., and JONES, C. J. (1994). A deductive object-oriented GIS for handling multiple representations. In T. C. Waugh, and R. G. Healey (eds.), *Advances in GIS Research.* Taylor & Francis, London, ii. 882–900.

KIM, W., and LOCHOVSKY, F. H. (eds.) (1989). *Object-Oriented Concepts, Databases and Applications.* Addison-Wesley, Reading, Mass.

KINEMAN, J. J. (1993). What is a scientific database. In M. F. Goodchild, L. T. Steyaert, B. O. Parks, (eds.), *Environmental Modeling and GIS.* Oxford University Press, New York, pp. 372–8.

KLEINER, A., and BRASSEL, K. E. (1986). Hierarchical grid structures for static geographic data bases. In M. Blakemore (ed.), *Proc. Autocarto London.* Imperial College, London, Sept.

KLIR, G. J., and FOLGER, T. A. (1988). *Fuzzy Sets, Uncertainty and Information.* Prentice Hall, Englewood Cliffs, NJ.

KOLLIAS, V. J., and VOLIOTIS, A. (1991). Fuzzy reasoning in the development of geographical information systems. FRSIS: a prototype soil information system with fuzzy retrieval capabilities. *International Journal of Geographical Information Systems*, 5: 209–24.

KOSKO, B. (1994). *Fuzzy Thinking: The New Science of Fuzzy Logic.* Harper Collins, London.

KUMAR, L., SKIDMORE, A. K., and KNOWLES, E. (1997). Modelling topographic variation in solar radiation in a GIS environment. *International Journal of Geographical Information Science*, 11: 475–98.

KUNIANSKY, E. L., and LOWTHER, R. A. (1993). Finite element mesh generation from mappable features. *International Journal of Geographical Information Systems*, 7: 395–405.

LAFLEN, J. M., and COLVIN, T. S. (1981). Effect of crop residue on soil loss from continuous row cropping. *Transactions of the American Society of Agricultural Engineers*, 24: 605–9.

LAGACHERIE, P., ANDRIEUX, P., and BOUZIGUES, R. (1996). Fuzziness and uncertainty of soil boundaries: from reality to coding in GIS. In P. A. Burrough and A. U. Frank (eds.), *Geographical Objects with Indeterminate Boundaries.* Taylor & Francis, London, pp. 275–86.

LAM, N. (1983). Spatial interpolation methods: a review. *American Cartographer*, 10: 129–49.

—— and DA COLA, L. (1993). *Fractals in Geography*, Prentice Hall, Englewood Cliffs, NJ.

LAMBECK, K. (1988). *Geophysical Geodesy.* Oxford Science Publications, Oxford, 718 pp.

LANGAS, S. (1997). Transboundary European GIS databases: a review of the Baltic Sea region experiences. In P. A. Burrough and F. I. Masser (eds.), *European Geographic Information Infrastructures: Opportunities and Pitfalls.* Taylor & Francis, London (in press).

LANGRAN, G. (1992). *Time in Geographic Information Systems.* Taylor & Francis, London.

LANTER, D. P., and VEREGIN, H. (1992). A research paradigm for propagating error in layer-based GIS. *Photogrammetric Engineering & Remote Sensing*, 58: 825–33.

LAPEDES, D. N. (ed.) (1976). *The McGraw-Hill Dictionary of Scientific and Technical Terms.* McGraw-Hill Book Company, New York, 1634 pp. + appendices.

LAPIDUS, D. F. (1987). *The Facts on File Dictionary of Geology and Geophysics.* Facts on File Publications, New York, 347 pp.

LASLETT, G. M., and MCBRATNEY, A. B. (1990*a*). Estimation and implications of instrumental drift, random measurement error and nugget variance of soil attributes— a case study for soil pH. *Journal of Soil Science*, 41: 451–72.

—— —— (1990*b*). Further comparison of several spatial prediction methods for soil pH. *Soil Sci. Am. J.* 54: 1553–8.

—— —— PAHL, P., and HUTCHINSON, M. (1987). Comparison of several spatial prediction methods for soil pH. *Journal of Soil Sciences*, 38: 325–41.

LAURINI, R., and PARIENTE, D. (1996). Towards a field-oriented language: first specifications. In P. A. Burrough and A. U. Frank (eds.), *Geographical Objects with Indeterminate Boundaries.* Taylor & Francis, London, 345 pp.

—— and THOMPSON, D. (1992). *Fundamentals of Spatial Information Systems.* Academic Press, London.

LAWRENCE, V. V. (1998). European Geoinformation Data Publishing: understanding the commercial volume industry. In P. A. Burrough and F. I. Masser, *European Geographic Information Infrastructures: Opportunities and Pitfalls.* Taylor and Francis, London.

LEAVESLEY, G. H., RESTROPO, P. J., STANNARD, L. G., FRANKOSKI, L. A., and SAUTINS, A. M. (1996). MMS: A Modeling Framework for Multidisciplinary Research and Operational applications. In M. F. Goodchild, L. T. Steyaert, B. O. Parks, C. Johnston, D. Maidment, M. Crane, and S. Glendinning (eds.), *GIS and Environmental Modeling: Progress and Research Issues.* GIS World Books, Fort Collins, Colo., pp. 155–8.

LEE, J. (1991*a*). Comparison of existing methods for building triangular irregular network models of terrain from grid digital elevation models. *International Journal of Geographical Information Systems*, 5: 267–85.

—— (1991*b*). Analysis of visibility sites on topographic surfaces. *International Journal of Geographical Information Systems*, 5: 413–29.

LEENAERS, H., BURROUGH, P. A., and OKX, J. P. (1989*a*). Efficient mapping of heavy metal pollution on floodplains by co-kriging from elevation data. In J. Raper (ed.), *Three Dimensional Applications in Geographic Information Systems.* Taylor & Francis, London, pp. 37–50.

—— OKX, J. P., and BURROUGH, P. A. (1989*b*). Co-kriging: an accurate & inexpensive means of mapping floodplain soil pollution by using elevation data. In M. Armstrong (ed.), *Geostatistics*, Proceedings of the third Geostatistics Congress, Avignon, Oct. 1988, Kluwer, pp. 371–82.

—— —— —— (1990). Comparison of spatial prediction methods for mapping floodplain soil pollution. *Catena*, 17: 535–50.

LEES, B. G. (1996*a*). Sampling strategies for machine learning using GIS. In M. F. Goodchild *et al., GIS and Environmental Modeling: Progress and Research Issues.* GIS World Books, Fort Collins, Colo., pp. 39–42.

—— (1996*b*). Neural network applications in the geosciences: an introduction. *Computers and Geosciences*, 22: 955–7.

LEGROS, J-P, KOLBL, O., and FALIPOU, P. (1996). Délimitation d'unités de paysage sur des photographies aériennes: éléments de réflexion pour la définition d'une méthode de tracé. *Étude et gestion des sols*, 3(2): 113–24.

LEVIALDI, S. (1980). Finding the edge. In J. C. Simon and R. M. Haralick (eds.), *Digital Image Processing.* Reidel, Dordrecht, pp. 105–48.

LILLESAND, T. M., and KIEFER, R. W. (1987). *Remote Sensing and Image Interpretation.* Wiley, New York.

LODWICK, W. A., MONSON, W., and SVOBODA, L. (1990). Attribute error and sensitivity analysis of map operations in geographical information systems. *International Journal of Geographical Information Systems*, 4: 413–27.

LORIE, R. A., and MEIER, A. (1984). Using a relational DBMS for geographical databases. *GeoProcessing*, 2: 243–57.

LUDER, P. (1980). *Das ökologische Ausgleichspotential der Landschaft.* Basler Beiträge zur Physiogeografie. Geografisches Institut der Universität Basel.

LYTLE, D. J., BLISS, N. B., and WALTMAN, S. W. (1996). Interpreting the State Soil Geographic Database (STATSGO). In: M. F. Goodchild, L. T. Steyaert, B. O. Parks, C. Johnston, D. Maidment, M. Crane, and S. Glendinning (eds.), *GIS and Environmental Modeling: Progress and Research Issues.* GIS World Books, Fort Collins, Colo., pp. 49–52.

MCALPINE, J. R., and COOK, B. G. (1971). Data reliability from Map Overlay. Proc. Australian and New Zealand Association for the Advancement of Science, 43rd Congress, Brisbane, May 1971, Section 21—Geographical Sciences.

MCBRATNEY, A. B. and DE GRUIJTER, J. J. (1992). A continuum approach to soil classification by modified fuzzy *k*-means with extragrades. *Journal of Soil Science*, 43: 159–76.

—— and WEBSTER, R. (1981). The design of optimal sampling schemes for local estimation & mapping of regionalized variables: 2 Program & examples. *Computers & Geosciences*, 7: 335–65.

—— —— (1983). Optimal interpolation & isarithmic mapping of soil properties: V. Co-regionalisation & multiple sampling strategy. *Journal of Soil Science*, 34: 137–62.

—— DE GRUIJTER, J. J., and BRUS, D. J. (1992). Spacial prediction and mapping of continuous soil classes. *Geoderma*, 54: 39–64.

MCCORMACK, J. E., GAHEGAN, M. N., ROBERTS, S. A., HOGG, J., and HOYLE, B. S. (1993). Feature-based derivation of drainage networks. *International Journal of Geographical Information Systems*, 7: 263–79.

MCDONALD, M. G., and HARBAUGH, A. W. (1988). Techniques of water resources investigations of the United States Geological Survey modular three-dimensional finite-difference groundwater flow model model. Reference Manual *Visual Modflow.* USGS, Washington, DC.

MCDONNELL. R. A. (1996). Including the spatial dimension: a review of the use of GIS in hydrology. *Progress in Physical Geography*, 21: 159–77.

—— and KEMP, K. K. (1995). *The International GIS Dictionary.* GeoInformation International, Cambridge.

MACDOUGAL, E. B. (1975). The accuracy of map overlays. *Landscape Planning*, 2: 23–30.

MCHARG, I. L. (1969). *Design with Nature.* Doubleday/Natural History Press, New York.

MACHOVER, C. (1989). *The C4 Handbook: CAD, CAM, CAE, CIM.* Tab Books Inc. Blue Ridge Summit, Pa., 438 pp.

MCLEISH, J. (1992). *Number. From ancient civilisations to the computer.* Flamingo, London, 266 pp.

MACMILLAN, R. A., FURLEY, P. A., and HEALEY, R. G. (1993). Using hydrological models and geographic information systems to assist with the management of surface water in agricultural landscapes. In R. Haines-Young, D. R. Green, and S. H. Cousins (eds.), *Landscape Ecology and GIS*, Taylor & Francis, London, pp. 181–209.

McRae, S. G., and Burnham, C. P. (1981). *Land evaluation.* Oxford University Press, Oxford.

Maguire, D. J., Goodchild, M. F., and Rhind, D. (eds.) (1991). *Geographical Information Systems: Principles and Applications.* Longman Scientific and Technical, Harlow, 2 vols.

Maidment, D. R. (1993). Developing a spatially distributed unit hydrograph by using GIS. In K. Kovar and H. P. Nachnebel (eds.), *Application of Geographic Information Systems in Hydrology and Water Resources Management HydroGIS 1993.* IAHS Publication No. 211. 181–92.

—— (1996*a*). GIS/Hydrologic Models of non-point source pollutants in the vadose zone. In D. L. Corwin and K. Loague (eds.), *Application of GIS to the Modeling of Non-Point Source Pollutants in the Vadose Zone.* Special SSSA Publication, Soil Science Society of America Inc., Madison, Wis., pp. 163–74.

—— (1996). Environmental modeling with GIS. In M. F. Goodchild, L. T. Steyaert, B. O. Parks, C. Johnston, D. Maidment, M. Crane, and S. Glendinning (eds.), *GIS and Environmental Modeling: Progress and Research Issues.* GIS World Books, Fort Collins, Colo., 315–24.

Makarovic, B. (1973). Progressive sampling for digital terrain models. *ITC Journal*, 1973(3): 397–416.

—— (1975), Amended strategy for progressive sampling. *ITC Journal* 1975(1): 117–28.

—— (1976). A digital terrain model system. *ITC Journal* 1976(1): 57–83.

—— (1977). Composite sampling for digital terrain models. *ITC Journal* 1977(3): 406–33.

Maling, D. H. (1973). *Coordinate Systems and Map Projections.* George Phillip, London.

—— (1992). *Coordinate Systems and Map Projections.* Pergamon, Oxford.

Mandelbrot, B. B. (1982). *The Fractal Geometry of Nature.* Freeman, New York.

Marble, D. F., Calkins, H., Peuquet, D. J., Brassel, K., and Wasilenko, M. (1981). *Computer Software for Spatial Data Handling* (3 vols.). Prepared by the International Geographical Union, Commission on Geographical Data Sensing and Processing for the US Department of the Interior, Geological Survey. IGU Ottawa, Canada, 1043 pp.

Mark, D. M., and Aronson, P. B. (1984). Scale-dependent fractal dimensions of topographic surfaces: an empirical investigation with application in geomorphology and computer mapping. *Math. Geol.* 16: 671–83.

—— and Lauzon, J. P. (1984). Linear quadtrees for Geographic Information Systems. *Proc. IGU Symposium on Spatial Data Handling, 20–4 Aug. 1984, Zurich,* pp. 412–31.

Marks, D., Dozier, J., and Frew, J. (1984). Automated basin delineation from digital elevation data. *GeoProcessing*, 2: 299–311.

Marsman, B., and de Gruijter, J. J. (1984). Dutch soil survey goes into quality control, in P. A. Burrough and S. W. Bie (eds.), *Soil Information Systems Technology.* PUDOC, Wageningen.

—— —— (1985). Quality of soil maps; a comparison of survey methods in a sandy area. *Soil Survey Papers* no. 15. Netherlands Soil Survey Institute, Wageningen.

Martin, D. (1996). An assessment of surface and zonal models of population. *International Journal of Geographical Information Systems*, 10: 973–89.

Martin, J. J. (1982). Organization of geographical data with quadtrees and least squares approximation. *Proc. Symposium on Pattern Recognition and Image Processing (PRIP).* Las Vegas, Nevada. IEEE Computer Society, pp. 458–65.

Masser, I., and Blakemore, M. (eds.) (1991). *Handling Geographical Information: Methodology and Potential Applications.* Longman Scientific and Technical, Harlow, 317 pp.

Mausbach, M., and Wilding, L. (1991). *Spatial Variabilities of Soils and Landforms.* Soil Science Society of American Special Publication No. 28, Madison, Wis.

Mead, D. A. (1982). Assessing data quality in geographic information systems. In C. J. Johannsen and J. L. Sanders (eds.), *Remote Sensing for Resource Management.* Soil Conservation Society of America, Ankeny, Iowa, pp. 51–62.

Meijerink, AMJ., de Brouwer, H. A. M., Mannaerts, C. M., and Valenzuela, C. R. (1994). *Introduction to the use of Geographic Information Systems for Practical Hydrology.* UNESCO International Hydrological Programme IHP-IV M2.3/International Institute for Aerospace Survey and Earth Sciences (ITC) Publication No. 23, 243 pp.

Meyer, L. D. (1981). How rain intensity affects interrill erosion. *Transactions of the American Society of Agricultural Engineers*, 24: 1472–5.

Miesch, A. T. (1975). Variograms and variance components in geochemistry and ore evaluation. *Geological Society of America Memoir*, 142: 333–40.

Middelkoop, H. (1997). *Embanked Floodplains in the Netherlands: Geomorphological Evolution over Various Time Scales.* Royal Netherlands Geographical Association/Faculty of Geographical Sciences, Utrecht University, Utrecht.

Miller, E. E. (1980). Similitude and scaling of soil-water phenomena. In D. Hillel (ed.), *Applications of Soil Physics*, Academic Press, New York.

Milne, P., Milton, S., and Smith , J. L. (1993). Geographical object-oriented databases: a case study. *International Journal of Geographical Information Systems*, 7: 39–55.

Mitasova, H., and Hofierka, J. (1993). Interpolation by regularized spline with tension: application to terrain modeling and surface geometry analysis. *Mathematical Geology*, 25: 657–69.

Mitasova, H., Mitas, L., Brown, W. M., Gerdes, D. P., Kosinovsky, I., and Baker, T. (1995). Modelling spatially and temporally distributed phenomena: new methods and tools for GRASS GIS. *International Journal of Geographical Information Systems*, 9: 433–46.

—— Hofierka, J., Zlocha, M., and Iverson, L. R. (1996). Modelling topographic potential for erosion and deposition using GIS. *International Journal of Geographical Information Systems*, 10: 629–42.

Monmonnier, M. (1993). *How to lie with maps.* University of Chicago Press, Chicago, 176 pp.

Moore, I. D. (1996). Hydrological Modeling and GIS. In M. F. Goodchild, L. T. Steyaert, B. O. Parks, C. Johnston, D. Maidment, M. Crane, and S. Glendinning (eds.), *GIS and Environmental Modeling: Progress and Research Issues.* GIS World Books, Fort Collins, Colo., 143–8.

—— Grayson, R. B., and Ladson, A. R. (1991). Digital terrain modeling: a review of hydrological, geomorphological and biological applications. *Hydrological Processes*, 5: 3–30.

—— Turner, A. K., Wilson, J. P., Jenson, S. and Band, L. (1993). GIS and land-surface-subsurface process modelling, in M. F. Goodchild, B.O. Parks, and L. T. Steyaert (eds.), *Environmental Modeling with GIS.* Oxford University Press, New York, pp. 196–230.

Morgan, R. P. C., Morgan, D. D. V., and Finney, H. J. (1984). A predictive model for the assessment of soil erosion risk. *Journal of Agricultural Engineering Research*, 30: 245–53.

Morris, G. D., and Heerdegen, R. G. (1988). Automatically derived catchment boundaries and channel networks and their hydrological applications. *Geomorphology*, 1: 131–41.

Morrison, J. L. (1971). Method-produced error in isarithmic mapping. Technical Monograph No. CA-5, American Congress on Surveying and Mapping, Washington, DC.

Mounsey, H. M. (1991). Multisource, multinational environmental GIS: lessons learnt from CORINE. In D. J. Maguire, M. F. Goodchild, and D. W. Rhind (eds.), *Geographical Information Systems, ii: Applications*, Longman Scientific and Technical, Harlow, pp. 185–200.

Nagy, G., and Wagle, S. G. (1979). Approximation of polygonal maps by cellular maps. *Comm. Ass. Comput. Mach.* 22: 518–25.

National Research Council (1994). *Promoting the National Spatial Data Infrastructure through Partnerships.* National Academy Press, Washington, DC.

Nichols, R. L., Looney, B. B., and Huddleston, J. E. (1992), 3-D digital imaging. *Environmental Science and Technology*, 26: 642–9.

Nielsen, D. R., and Bouma, J. (1985). *Spatial Analysis of Soil Data.* PUDOC, Wageningen.

Norbeck, S., and Rystedt, B. (1972). *Computer Cartography.* Studentlitteratur Lund, University of Lund, 315 pp.

Nortcliff, S. (1978). Soil variability and reconnaissance soil mapping: a statistical study. *Norfolk. J. Soil Science*, 29: 403–18.

Odeh, I. O. A., McBratney, A. B., and Chittleborough, D. J. (1990). Design of optimal sample spacings for mapping soil using fuzzy-k-means and regionalized variable theory. *Geoderma*, 47: 93–122.

Okx, J. P., and Kuipers, B. R. (1991). 3-Dimensional probability mapping for soil remediation purposes. In J.-J. Harts, H. F. L. Ottens, and H. J. Scholten (eds.), *Proc. EGIS'91.* EGIS Foundation, Utrecht, 765–74.

Olea, R. A. (1991). *Geostatistical Glossary and Multilingual Dictionary.* International Association for Mathematical Geology Studies in Mathematical Geology No. 3, Oxford University Press, New York, 177 pp.

Openshaw, S. (1977). A geographical solution to scale and aggregation problems in region-guilding, partitioning, and spatial modelling. *Institute of British Geographers Transactions*, 2 (ns): 459–72.

—— and Taylor, P. J. (1979). A million or so correlation coefficients: Three experiments on the modifiable area problem. In N. Wrigley (ed.). *Statistical Methods in the Spatial Sciences*, Pion, London, pp. 127–44.

Oswald, H., and Raetzsch, H. (1984). A system for generations and display of digital elevation models. *Geo-Processing*, 2: 197–218.

Ozemoy, V. M., Smith, D. R., and Sicherman, A. (1981). Evaluating computerized geographic information systems using decision analysis. *Interfaces*, 11: 92–8.

Pannatier, Y. (1996). *Variowin: Software for Spatial Data Analysis in 2D.* Statistics and Computing, Springer Verlag, Berlin, 91 pp.

Parker, H. D. (1988). The unique qualities of a geographic information system: a commentary. *Photogrammetric Engineering and Remote Sensing*, 54: 1547–9.

Parratt, L. G. (1961), *Probability and Experimental Errors.* Wiley, New York, 255 pp.

Pavlidis, T. (1982). *Algorithms for Graphics and Image Processing.* Springer Verlag, Berlin.

Pebesma, E. (1996), *Mapping Groundwater Quality in the Netherlands.* Royal Dutch Geographical Association/ Faculty of Geographical Sciences, Utrecht University.

—— and Wesseling, C. G. (1998), GSTAT, a program for geostatistical modelling, prediction and simulation. *Computers and Geosciences*, 24: 17–31.

Perkal, J. (1966). On the length of empirical curves. Discussion paper 10, Ann Arbor Michigan Inter-University Community of Mathematical Geographers.

Peuker, T. K. (now Poiker) and Chrisman, N. 1975. Cartographic Data Structures. *The American Cartographer*, 2(1), 55–69.

—— Fowler, R. J., Little, J. J., and Mark, D. M. (1978). The triangulated irregular network. In *Proc. of the DTM Symposium, American Society of Photogrammetry—American Congress on Survey and Mapping,* St Louis, Missouri, pp. 24–31.

Peuquet, D. J. (1977). Raster data handling in geographic information systems. In G. Dutton (ed.), *Proc. 1st Int. Advanced Study Symposium on Topological Data Structures.* Laboratory for Computer Graphics and Spatial Analysis, Harvard.

—— (1979). Raster processing: an alternative approach to automated cartographic data handling. *Am. Cartogr.* 6: 129–39.

—— (1984*a*). A conceptual framework and comparison of spatial data models. *Cartographica*, 21: 66.

—— (1984*b*). Data structures for a knowledge-based geographic information system. In *Proceedings of the International Symposium on Spatial Data Handling.* 20–4 Aug. 1984, Zurich, pp. 372–91.

—— and Marble, D. F., (eds.) (1990). *Introductory Readings in Geographic Information Systems.* Taylor & Francis, London, 371 pp.

Piwowar, J. M., LeDrew, E. F., and Dudycha, D. J. (1990). Integration of spatial data in vector and raster formats in geographical information systems. *International Journal of Geographical Information Systems*, 4: 429–44.

Poggio, T. (1984). Vision by man and machine. *Scientific American*, 250(4): 68–78.

Poiker, T. K. (formerly Peuker) (1982). Looking at computer cartography. *GeoJournal*, 6: 241–9.

Powell, B., McBratney, A. B., and MacLeod, D. A. (1991). The application of fuzzy classification to soil pH profiles in the Lockyer Valley, Queensland, Australia. *Catena*, 18: 409–20.

Quinn, P., Beven, K., Chevalier, P., and Planchon, O. (1991). The prediction of hillslope flow paths for distributed hydrological modelling using digital terrain models. *Hydrological Processes*, 5: 59–79.

Quiroga, C. A., Singh, V. P., and Iyengar, S. S. (1996). Spatial data characteristics. In V. P. Singh, and M. Fiorentino (eds.), *Geographical Information Systems in Hydrology* Kluwer Academic, Dordrecht, pp. 65–89.

Raper, J., and Livingstone, D. (1995). Development of a geomorphological spatial model using object-oriented design. *International Journal of Geographical Information Systems*, 9: 359–84.

—— Rhind, D., and Shepherd, J. (1992). *Postcodes: The New Geography.* Longman Scientific and Technical, Harlow, 322 pp.

Rengess, N. (ed.) (1994). *Engineering Geology of Quaternary Sediments*, A. A. Balkema, Rotterdam.

Reumann, K., and Witkam, A. P. M. (1974). Optimizing curve segmentation in computer graphics. *International Computing Symposium 1973.* North-Holland, Amsterdam, pp. 467–72.

Rhind, D. (1977). Computer-aided cartography. *Trans. Inst. British Geographers,* NS 2: 71–96.

—— and Green, N. P. A. (1988). Design of a geographical information system for a heterogeneous scientific community. *International Journal of Geographical Information Systems*, 2: 171–89.

Ripley, B. D. (1981). *Spatial Statistics*, Wiley, New York.

Ritsema Van Eck, J. R. (1993). *Analyse van transport netwerken in GIS vor sociaal-geografisch onderzoek.* Netherlands Geographical Studies 164, Koninklijk Nederlands Aardrijkskundige Genootschap, University of Utrecht, 195 pp.

Ritter, P. (1987). A vector-based slope and aspect generation algorithm. *Photogrammetric Engineering and Remote Sensing*, 53: 1109–11.

RIVM (1994). *The Preparation of a European Land Use Database.* RIVM Report no. 712401001, Rijksinstituut voor Volksgezondheid en Milieuhygiene, Bilthoven, the Netherlands.

Rivoirard, J. (1994). *Introduction to Disjunctive Kriging and Non-linear Geostatistics.* Clarendon Press, Oxford.

Robinove, C. J. (1986). *Principles of Logic and the Use of Digital Geographic Information Systems.* US Geological Survey Circular 977, US Geological Survey, Denver, Colo., 19 pp.

Robinson, A. H., Sale, R., and Morrison, J. (1978). *Elements of Cartography* (4th edn.). Wiley, New York.

Rose, E. (1975). Erosion et ruissellment en Afrique de l'ouest: vingt années de mesures en petites parcelles experimentales. Cyclo. ORSTOM. Abidjan, Ivory Coast.

Rosenfeld, A. (1980). Tree structures for region representation. In H. Freeman and G. G. Pieroni (eds.), *Map Data Processing.* Academic Press, New York, pp. 137–50.

—— and Kak, A. (1976). *Digital Picture Processing.* Academic Press, New York.

Rossiter, D. G. (1996). A theoretical framework for land evaluation. *Geoderma*, 72: 165–90.

Roubens, M. (1982). Fuzzy clustering algorithms and their cluster validity. *Eur. J. Opt. Res.*, 10: 294–301.

Salgé, F. (1997). From an understanding of the European GI Economic activity to the reality of a European datasef. In Burrough and Masser 1997.

Salomé, A. I., van Dorsser, H. J., and Rieff, P. L. (1982). A comparison of geomorphological mapping systems. *ITC Journal*, 1982(3): 272–4.

Salton, G., and McGill, M. J. (1983). *Introduction to Modern Information Retrieval.* McGraw-Hill, New York, 448 pp.

Samet, H. (1990*a*). *Applications of Spatial Data Structures: Computer Graphics, Image Processing and Geographical Information Systems*. Addison-Wesley, Reading, Mass.

—— (1990*b*). *The Design and Analysis of Spatial Data Structures*. Addison-Wesley, Reading, Mass.

Scheidegger, A. E. (1970). Stochastic models in hydrology. *Water Resources Research*, 6: 750–5.

Schell, D. (1995*a*). What is the meaning of standards consortia? *GIS World* (Aug.) 82.

—— (1995*b*). Harnessing change. Editorial in *OpenGIS*, Newsletter included in *Geo Info Systems*, May.

Scott, J. S. (1980). *The Penguin Dictionary of Civil Engineering*. Penguin Books, Harmondsworth, 308 pp.

Serra, J. (1968). Les Structures gigognes: morphologie mathématique et interprétation métallogénique. *Mineral. Deposita (Berl.)*, 3: 135–54.

Sharpnack, D. A., and Atkins, G. (1969). An algorithm for computing slope and aspect from elevations. *Photogrammetric Engineering and Remote Sensing*, 35: 247–8.

Sheehan, D. E. (1979). A discussion of the SYMAP program. Harvard Library of Computer Graphics, *Mapping Collection, ii: Mapping Software and Cartographic Databases*, Cambridge, Mass., pp. 167–79.

Skidmore, A. K. (1989). A comparison of techniques for calculating gradient and aspect from a gridded digital elevation model. *International Journal of Geographical Information Systems*, 3: 323–34.

Smedley, B., and Aldred, B. K. (1980). Problems with Geodata, in A. Blaser (ed.), *Data Base Techniques for Pictorial Applications*. Springer-Verlag, Berlin, pp. 539–54.

Smith, D. D., and Wischmeier, W. H. (1957). Factors affecting sheet and rill erosion. *Trans. Am. Geophys. Union*, 38: 889–96.

Smith, T. R., Menon, S., Starr, J. L., and Estes, J. E. (1987). Requirements and principles for the implementation and construction of large-scale geographic information systems. *International Journal of Geographical Information Systems*, 1: 13–31.

Sneath, P. H. A., and Sokal, R. R. (1973). *Numerical Taxonomy*, Freeman, San Francisco, 573 pp.

Soil Survey Staff (1951). *Soil Survey Manual*. USDA Handbook No. 18, US Govt. Printing Office, Washington, DC.

—— (1976). *Soil Taxonomy*. US Govt. Printing Office, Washington, DC.

Sokolnikoff, I. S., and Sokolnikoff, E. S. (1941). *Higher Mathematics for Engineers and Physicists*. McGraw-Hill, New York, 587 pp.

Solow, A. R. (1986). Mapping by simple indicator kriging. *Mathematical Geology*, 18: 335–51.

—— & Gorelick, S. M. (1986). Estimating monthly streamflow values by co-kriging. *Mathematical Geology*, 18: 785–809.

Steiner, D., and Matt, O. F. (1972). *Computer Program for the Production of Shaded Choropleth and Isarithmic Maps on a Line Printer*. User's manual. Waterloo, Ont.

Steinitz, C., and Brown, H. J. (1981). A computer modelling approach to managing urban expansion. *Geoprocessing*, 1: 341–75.

Stevens, P. (ed.) (1988). *Oil and Gas Dictionary*. Macmillan Reference Books, London, 270 pp.

Stienstra, P., and Van Deen, J. K. (1994). Field data collection techniques: unconventional sounding and sampling. In N. Rengers (ed.), *Engineering Geology of Quaternary Sediments*. Balkema, Rotterdam, pp. 41–56.

Stocking, M. (1981). A working model for the estimation of soil loss suitable for underdeveloped areas. Development Studies Occasional Paper No. 15, University of East Anglia, UK.

Switzer, P. (1975). Estimation of the accuracy of qualitative maps. In J. C. Davis and M. J. McCullagh (eds.), *Display and Analysis of Spatial Data*. Wiley, New York, 380 pp.

Takeyama, T., and Couclelis, H. (1997). Map dynamics: integrating cellular automata and GIS through GEO-Algebra. *International Journal of Geographical Information Science*, 11: 73–92.

Tang, H. J. and van Ranst, E. (1992). Testing of fuzzy set theory in land suitability assessment for rainfed grain maize production. *Pedologie*, 42(2): 129–47.

—— Debaveye, J., Ruan, D. and van Ranst, E. (1991). Land suitability classification based on fuzzy set theory. *Pedologie*, 41(3): 277–90.

Taylor, C. C., and Burrough, P. A. (1986). Multiscale sources of spatial variation in soil III. Improved methods for fitting the nested model to one-dimensional semivariograms. *Mathematical Geology*, 18: 811–21.

Taylor, J. R. (1982). *An Introduction to Error Analysis*. University Science Books, Oxford University Press, Oxford.

Teicholz, E., and Berry, B. J. L. (1983). *Computer Graphics and Environmental Planning*. Prentice Hall, Englewood Cliffs, NJ.

Ter Steeg, H., Jetten, V., Polak, M., and Werger, M. (1993). Tropical rainforest types and soils of a watershed in Guyana, South America. *Journal of Vegetation Science*, 4: 705–16.

Thacher, P. (1996). 'Sustainable Development' needs GIS. Unpublished keynote address, Tenth Annual GIS-World Symposium on Geographic Information Systems, Vancouver, British Columbia, Canada, 20 Mar. 1996.

Thapa, K., and Bossler, J. (1992). Accuracy of spatial data used in Geographic Information Systems. *Photogrammetric Engineering and Remote Sensing*, 58: 841–58.

Tobler, W. (1979). Smooth pycnophylactic interpolation for geographic regions. *J. American Statistical Association*, 74(367): 519–36.

—— (1995). The resel-based GIS. *International Journal of Geographical Information Systems*, 9: 95–100.

—— Deichmann, U., Gottsegen, J., and Maloy, K. (1995). The Global Demography Project. *NCGIA Technical Report* 95–6. National Center for Geographic Information and Analysis, University of California, Santa Barbara, 75 pp.

Tomlin, C. D. (1983) A map algebra. In *Proc. Harvard Computer Conf, 1983, 31 July – 4 August, Cambridge, Mass.*

—— (1990). *Geographic Information Systems and Cartographic Modeling*. Prentice Hall, Englewood Cliffs, NJ, 249 pp.

Tomlinson, R. F. (1984). Geographic Information Systems: a new frontier. *Proceedings of the International Symposium on Spatial Data Handling*. 20–4, Aug. 1984, Zurich, pp. 1–14.

—— and Boyle, A. R. (1981). The state of development of systems for handling natural resources inventory data. *Cartographica*, 18: 65–95.

—— Calkins, H. W. and Marble, D. F. (1976). *Computer Handling of Geographic Data*. UNESCO, Geneva.

Travis, M. R., Elsner, G. H., Iverson, W. D., and Johnson, C. G. (1975). VIEWIT computation of seen areas, slope and aspect for land use planning. US Department of Agriculture Forest Service General Technical Report PSW 11/1975, Pacific Southwest Forest and Range Experimental Station, Berkeley, Calif.

Trier, O. D., and Taxt, T. (1994). Evaluation of binarization methods for utility map images. Proceedings ICIP-94, Austin Texas, vol ii. 1046–50.

Tucker, C. J. (1979). Red and photographic infrared linear combinations for monitoring vegetation. *Remote Sensing of Environment*, 8: 127–50.

Ullman, J. D. (1980). *Principles of database systems*. Computer Science Press, Potomac, Md.

UNEP/FAO (1994). Report of the UNEP/FAO Expert Meeting on harmonizing land cover and land use elassifications. GEMS Report series No. 25, UNEP/FAO Nairobi.

van Deursen, W. P. A. (1995). *Geographical Information Systems and Dynamic Models*. Ph.D. thesis, Utrecht University, NGS Publication 190, 198 pp.

—— and Burrough, P. A. (1998). *Dynamic modelling with GIS*. Oxford (forthcoming).

van Oosterom, P. (1993). *Reactive Data Structures for Geographic Information Systems*. Oxford University Press, Oxford.

van Reeuwijk, P. (1982). Laboratory methods and data quality. Program for soil characterization: a report on the pilot round. Part I. CEC and texture. *Proc. 5th Int. Classification Workshop*, Khartoum, Sudan, Nov. 1982, 58 pp.

—— (1984). Idem: Part II. Exchangeable bases, base saturation and pH. International Soil and Reference Centre, Wageningen, 28 pp.

van Roessel, J. W., and Fosnight, E. A. (1984). A relational approach to vector data structure conversion. In *Proc. IGU Int. Symp. on Spatial Data Handling*, 20–4 Aug., Zurich. pp. 78–95.

Varekamp, C., Skidmore, A. K., and Burrough, P. A. (1996). Spatial interpolation using public domain software. *Photogrammetric Engineering and Remote Sensing*, 62: 845–54.

Vink, A. P. A. (1963). Planning of soil surveys in land development. *Pub. Int. Inst. Land Reclamation and Improvement* 10. 50 pp.

—— (1981). *Landschapsecologie en landgebruik*. Scheltema en Holkema, Amsterdam.

Voltz, M., and Webster, R. (1990). A comparison of kriging, cubic splines and classification for predicting soil properties from sample information. *Journal of Soil Science*, 41: 473–90.

Voss, R. (1985). Random fractal forgeries: from mountains to music. In S. Nash (ed.), *Science and Uncertainty*, IBM UK Ltd/Science Reviews, Northwood, England, pp. 69–88.

Vriend, S. P., van Gaans, P. F. M., Middelburg, J., and de Nijs, A. (1988). The application of fuzzy *c*-means cluster analysis and non-linear mapping to geochemical datasets: examples from Portugal. *Appl. Geochem.*, 3: 213–24.

Walling, D. E. and Bradley, S. B. (1988). Transport and redistribution of Chernobyl fallout radionuclides by fluvial processes: some preliminary evidence. *Environmental Geochemistry and Health*, 10: 35–9.

—— Rowan, J. S., and Bradley, S. B. (1989). Sediment-associated transport and redistribution of Chernobyl fallout radionuclides. *Sediment and the Environment* (Proc. Baltimore Symposium 1989), IAHS Publication no. 184, pp. 37–45.

Wang, F., Hall, G. B., and Subaryono (1990). Fuzzy information representation and processing in conventional GIS software: database design and application. *International Journal of Geographical Information Systems*, 4: 261–83.

Webber, R. (1997). Developments in cross-border standards for Geodemographic segmentation. In Burrough and Masser 1997 (in press).

Weber, W. (1978). Three types of map data structures, their ANDs and NOTs, and a possible OR. G. Dutton (ed.), *Harvard Papers on Geographic Information Systems: First International Advanced Study Symposium on Topological Data Structures for Geographic Information Systems*, vol. iv. Laboratory for Computer Graphics and Spatial Analysis, Graduate School of Design, Harvard University.

Webster, R. (1968). Fundamental objections to the 7th approximation. *Journal of Soil Science*, 19: 354–66.

—— (1973). Automatic soil boundary location from transect data. *Mathematical Geology*, 5: 27–37.

WEBSTER, R. (1977). *Quantitative and Numerical Methods in Soil Classification and Survey*. Oxford University Press, Oxford, 269 pp.

—— (1978). Optimally partitioning soil transects. *Journal of Soil Science*, 29: 388–402.

—— (1980). DIVIDE: A FORTRAN IV program for segmenting multivariate one-dimensional spatial series. *Computers and Geosciences*, 6: 61–8.

—— and BECKETT, P. H. T. (1970). Terrain classification and evaluation using air photography: a review of recent work at Oxford. *Photogrammetrica*, 26: 51–75.

—— and BURGESS, T. M. (1984). Sampling and bulking strategies for estimating soil properties in small regions. *Journal of Soil Science*, 35: 127–40.

—— and MCBRATNEY, A. B. (1987). Mapping soil fertility at Broom's Barn by simple kriging. *J.Sci.Food Agric.* 38: 97–115.

Webster, R., and OLIVER, M. A. (1990). *Statistical Methods in Soil & Land Resources Survey*, Oxford University Press, Oxford.

—— and WONG, I. F. T. (1969). A numerical procedure for testing soil boundaries interpreted from air photographs. *Photogrammetria*, 24: 59–72.

WEIBEL, R. *et al.* (1998). *Digital Elevation Models*. Oxford University Press, Oxford (in press).

WEIDEN, M. J. J. VAN DER, VAN GAANS, P. F. M., and VRIEND, S. P. (1992). Characterization of contaminated harbour soil with statistical tools. Draft proceedings First Conference of the Working Group on Pedometrics of the International Society of Soil Science, Pedometrics-92: Developments in Spatial Statistics for Soil Science, pp. 191–209.

WEIZENBAUM, J. (1976). *Computer Power and Human Reason*. W. H. Freeman, San Francisco.

WEL, F. J. M. VAN DER, and HOOTSMANS, R. M. (1993). Visualisation of quality information as an indispensible part of optimal information extraction from a GIS. *Proc. 16th International Cartographic Conference*, Cologne, 2–9 May 1993, pp. 881–97.

WELLAR, B., and WILSON, P. (1993). Contributions of GIS concepts and capabilities to scientific enquiry: initial findings. *GIS/LIS Proceedings 1993*, pp. 753–67.

—— —— (1995). Impact of GIS on spatial theorizing: a status and progress report. *Proceedings GIS/US '95*, ii. 1026–36.

WENTWORTH, C. K. (1930). A simplified method for determining the average slope of land surfaces. *Am. J. Sci. Series*, 5, 20(117): 184–94.

WESSELING, C. G., and HEUVELINK, G. B. M. (1991). Semi-automatic evaluation of error propagation in GIS operations. In J. Harts *et al.* (eds.), *Proceedings EGIS '91* (Utrecht: EGIS Foundation), 1228–37.

—— KARSSENBERG, D., BURROUGH, P. A., and VAN DEURSEN, W. P. A. (1996). Integrating dynamic environmental models in GIS: the development of a Dynamic Modelling language. *Transactions in GIS*, 1: 40–8.

WHITTEN, D. G. A., and BROOKS, J. R. V. (1972). *The Penguin Dictionary of Geology*. Penguin Books, Harmondsworth, 511 pp.

WHITTINGTON, R. P. (1988). *Database Systems Engineering*. Oxford University Press, Oxford.

WHITTOW, J. (1984). *The Penguin Dictionary of Physical Geography*. Penguin Books, Harmondsworth, 591 pp.

WIELEMAKER, W. G., and BOXEM, H. W. (1982). Soils of the Kisii area, Kenya. Agricultural Research Report 922, PUDOC/Agricultural University, Wageningen.

WILDING, L. P., JONES, R. B., and SCHAFER, G. M. (1965). Variation of soil morphological properties within Miami, Celina and Crosby mapping units in West-Central Ohio. *Proc. Soil Sci. Soc. Am.* 29(6): 711–17.

WIRTH, N. (1976). *Algorithms + Data = Structures*. Prentice-Hall, Englewood Cliffs, NJ.

WISCHMEIER, W. H., and SMITH, D. D. (1978). *Predicting Rainfall Erosion Losses*. Agricultural Handbook 537, USDA, Washington DC.

WORBOYS, M. F. (1994). Object oriented approaches to geo-referenced information, *International Journal of Geographical Information Systems*, 8: 385–99.

—— (1995). *GIS—a Computing Perspective*. Taylor & Francis, London.

—— HEARNSHAW, H. M., and MAGUIRE, D. J. (1990). Object-oriented data modelling for spatial databases. *International Journal of Geographical Information Systems*, 4: 369–83.

YAGER, R. R., and FILEV, D. P. (1994). *Essentials of Fuzzy Modeling and Control*. John Wiley, New York.

YEARSLEY, C. M., and WORBOYS, M. F. (1995). A deductive model of planar spatio-temporal objects. In P. Fisher (ed.), *Innovations in GIS* 2. Taylor & Francis, London, pp. 43 ff.

YOELI, P. (1982). Ueber digitale Geländemodelle und deren computergestützte kartographische und kartometrische Auswertung. *Vermess. Photogramm. Kulturtech.* 2/82: 34–9.

—— (1984), Cartographic contouring with computer and plotter. *American Cartographer*, 11: 139–55.

ZADEH, L. A. (1965). Fuzzy sets. *Information and Control*, 8(3): 338–53.

ZEVENBERGEN, L. W., and THORNE, C. R. (1987). Quantitative analysis of land surface topography. *Earth Surface Processes and Landforms*, 12: 47–56.

ZONNEVELD, J. I. S. (1973). Convergentie en luchtfoto-interpretatie. *K.N.A.G. Geografisch Tijdschrift*, 7(1): 38–48.

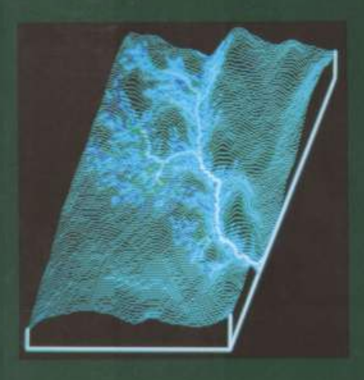
Index

This new edition of the best-selling *Principles of Geographical Information Systems for Land Resources Assessment* has been completely revised and brought up to date.



This classic text now features:

- Essential theoretical background to concepts used in spatial information handling
- Fundamentals of coding spatial data and storing them in electronic databases, including linking data from ground survey and remote sensing
- Point sampling and interpolation, geostatistics and digital elevation models
- New examples including site location analysis, marketing, route optimizing, land degradation, and the distribution of Chernobyl radionuclides
- Special attention for data quality and for the principles and use of fuzzy logic
- Analysis of world-wide spatial data provision including the Internet
- Summaries and questions for students and teachers
- Over 180 illustrations including line-drawings, half-tones, and a colour section.



Geographical information systems (GIS) are increasingly essential to students and professionals in all areas of the geographical and environmental sciences. Spatial data and GIS packages are currently used to solve many practical problems ranging from issues in consumer retailing to land degradation, hydrology and landscape ecology.

Using real world examples this new book provides a comprehensive yet concise introduction to the theory and practice of GIS for undergraduates and professionals alike in disciplines ranging from hydrology to epidemiology, from spatial planning to agriculture.

'everything you need to know about the evolution of GIS in its third decade . . . my recommended course text for many years to come.'
Professor David Unwin, Birkbeck College, University of London

'this is a book written by two scientists who use GIS every day . . . That makes it an ideal text for courses in environmental science for students who need to know the principles on which GIS operates'
Professor Michael Goodchild, University of California, Santa Barbara

Peter A. Burrough is Professor of Physical Geography at Utrecht University
Rachael A. McDonnell is a lecturer at Hertford College, Oxford

OXFORD UNIVERSITY PRESS

